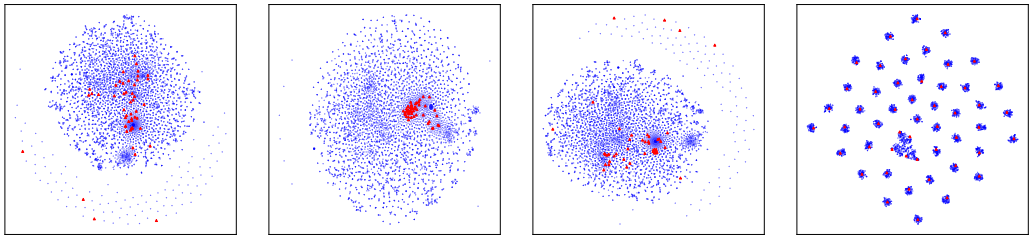
NEURAL TOPIC MODELING WITH EMBEDDING CLUSTERING REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Topic models have been prevalent for decades with various applications. However, existing topic models severely suffer from topic collapsing: the discovered topics semantically collapse towards each other, leading to highly repetitive topics, insufficient topic discovery, and damaged model interpretability. In this paper, we propose the Embedding Clustering Regularization Topic Model (ECRTM) to solve this collapsing problem. Specifically, we propose a novel embedding clustering regularization which forms distinct word embedding clusters with topic embeddings as centers and avoids trivial clustering solutions by producing sparse soft-assignments and fulfilling cluster size constraints. Through jointly optimizing the neural topic modeling objective and the regularization objective, ECRTM learns diverse and coherent topics together with high-quality topic distributions of documents at the same time. Extensive experiments on benchmark datasets demonstrate our ECRTM effectively alleviates the topic collapsing issue and consistently surpasses state-of-the-art baselines with substantially improved performance.



(a) ETM (b) NSTM (c) WeTe (d) **ECRTM**

Figure 1: t-SNE visualization (van der Maaten & Hinton, 2008) of word embeddings (●) and topic embeddings (▲) under 50 topics. This shows while the topic embeddings (▲) all collapse together in state-of-the-art models ETM (Dieng et al., 2020), NSTM (Zhao et al., 2021b), and WeTe (Wang et al., 2022), the topic embeddings (▲) of our proposed **ECRTM** are away from each other and are the centers of distinct word embedding clusters (●).

1 INTRODUCTION

Topic models have achieved great success in document analysis with various downstream applications (Boyd-Graber et al., 2017) like content recommendation (McAuley & Leskovec, 2013), summarization (Ma et al., 2012), and information retrieval (Wang et al., 2007). Current topic models can be roughly classified as two lines: conventional topic models based on probabilistic graphical models (Blei et al., 2003) or

Topic#1:	just show even come time one good really going know
Topic#2:	just even really something come going like actually things get
Topic#3:	just one even something come way really like always good
Topic#4:	just get going come one know even really something way
Topic#5:	just like inside get even look come one everything away

Table 1: Illustration of topic collapsing. Top related words of the discovered topics by NSTM (Zhao et al., 2021a) on IMDB. These topics are uninformative and have many repetitive words that are underlined.

matrix factorization (Kim et al., 2015; Shi et al., 2018) and neural topic models (Miao et al., 2016; 2017; Srivastava & Sutton, 2017).

However, despite current achievements, existing topic models commonly suffer from *topic collapsing*, namely the discovered topics tend to semantically collapse towards each other (Srivastava & Sutton, 2017), as exemplified in Table 1. This brings about several issues: (**1**) Topic collapsing results in highly repetitive topics that are less useful for downstream tasks (Nan et al., 2019). Table 1 shows these topics are uninformative and trivial. (**2**) Topic collapsing incurs insufficient topic discovery. Many other topics are undisclosed from the document collection, and thus the discovered topics are not inclusive enough. (**3**) Topic collapsing damages the interpretability of topic models. It becomes highly difficult to understand the real underlying topics that a document contains (Huynh et al., 2020; Dieng et al., 2020). In consequence, topic collapsing impedes the utilization and extension of topic models, and therefore it is crucial to overcome this challenge for reliable topic modeling applications.

To address the above problems, we in this paper propose a novel neural topic model, named **Embedding Clustering Regularization Topic Model** (**ECRTM**). First, we investigate the reason for the topic collapsing. Figure 1 shows the topic embeddings mostly collapse together in the state-of-the-art methods including the newest NSTM (Zhao et al., 2021b) and WeTe (Wang et al., 2022), which produces similar topic-word distributions and thus results in topic collapsing. Then to avoid the collapsing of topic embeddings, we propose the novel Embedding Clustering Regularization (ECR). It solves a devised clustering problem that requires topic embeddings to be the centers of distinct word embedding clusters and be away from each other instead of collapsing together. Different from previous clustering methods, ECR converts this clustering problem into a specifically defined optimal transport problem of the topic and word embeddings. As such, ECR can support joint optimization, produce sparse soft-assignments, and fulfill cluster size constraints, which enables it to form effective clusters and avoid trivial clustering solutions. As a result, ECR avoids the collapsing of topic embeddings by encouraging topic embeddings to be away from each other and be the centers of distinct word embedding clusters. Based on ECR, our proposed ECRTM alleviates the topic collapsing issue and further substantially improves topic modeling performance. Figure 1 shows the effectiveness of our ECRTM. We conclude the main contributions of our paper as follows:

- We propose a novel embedding clustering regularization that can form distinct word embedding clusters with topic embeddings as centers and avoid trivial clustering solutions, which effectively alleviates the topic collapsing issue.
- We further propose a new neural topic model that jointly optimizes the topic modeling objective and the embedding clustering regularization objective. Our model can learn diverse and coherent topics together with high-quality doc-topic distributions at the same time.
- We conduct extensive experiments on benchmark datasets and demonstrate that our model significantly mitigates the topic collapsing issue and surpasses state-of-the-art baseline methods with substantially improved topic modeling performance.

2 Related Work

Conventional Topic Models Conventional topic models (Hofmann, 1999; Blei et al., 2003; Blei & Lafferty, 2006; Mimno et al., 2009; Boyd-Graber et al., 2017) commonly employ probabilistic graphical models to model the generation process of documents with topics as latent variables. They infer model parameters with MCMC methods like Gibbs Sampling (Steyvers & Griffiths, 2007) or Variational Inference (Blei et al., 2017). Another line of work uses matrix factorization to model topics (Yan et al., 2013; Kim et al., 2015; Shi et al., 2018). Conventional topic models usually need model-specific derivations for different modeling assumptions.

Neural Topic Models Due to the success of Variational AutoEncoder (VAE, Kingma & Welling, 2014; Rezende et al., 2014), several neural topic models have been proposed (Miao et al., 2016; 2017; Srivastava & Sutton, 2017; Burkhardt & Kramer, 2019; Wu et al., 2020; Dieng et al., 2020; Meng et al., 2020). Neural topic models commonly optimize the KL divergence and the reconstruction error between input and output by directly applying gradient backpropagation. Some studies directly cluster pre-trained word or sentence embeddings to produce topics (Sia et al., 2020; Zhang

et al., 2022), but they are *not* topic models since they cannot infer the topic distributions of documents as required. Recent NSTM (Zhao et al., 2021b) and WeTe (Wang et al., 2022) measure the reconstruction error with optimal transport and conditional transport distance. However, they still suffer from topic collapsing, producing repetitive and useless topics as shown in our experiments (Sec. 4.2). Different from these, our proposed model aims to address the topic collapsing issue and substantially improve topic modeling performance. Instead of mainly minimizing reconstruction error as existing methods, our model proposes a novel embedding clustering regularization to form effective word embedding clusters with topic embeddings as centers. Then our model learns topics under this effective regularization, which particularly alleviates topic collapsing.

3 METHODOLOGY

3.1 PRELIMINARY OF TOPIC MODELING

In this section, we recall the preliminary background of topic modeling following LDA (Blei et al., 2003). Consider a document collection $\mathbf{X}$ with V unique words (vocabulary size), and each document is denoted as $\mathbf{x}$. We require to discover K topics from this document collection. The k-th latent topic is defined as a distribution over all words (topic-word distribution), denoted as $\boldsymbol{\beta}_k \in \mathbb{R}^V$. We have $\boldsymbol{\beta}=(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) \in \mathbb{R}^{V \times K}$ as the topic-word distribution matrix. The topic distribution of a document (doc-topic distribution) refers to what topics it contains, noted as $\boldsymbol{\theta} \in \Delta_K$. [1] LDA-like models mostly infer parameters with Gibbs Sampling (Steyvers & Griffiths, 2007) or Variational Inference (Blei et al., 2017). Recent studies apply neural variational inference (Kingma & Welling, 2014) to learn topics (Miao et al., 2016; Srivastava & Sutton, 2017; Dieng et al., 2020; Zhao et al., 2021b), which can directly apply gradient backpropagation.

3.2 WHAT CAUSES TOPIC COLLAPSING?

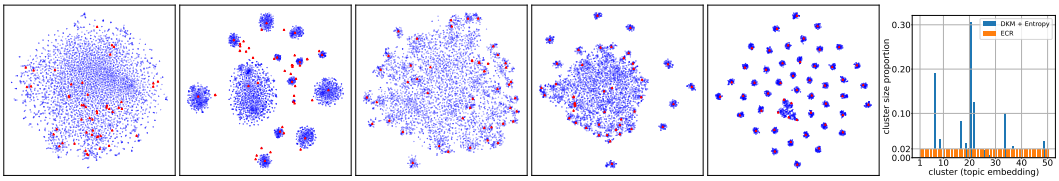
Despite the achievements of topic modeling, we find existing methods mostly suffer from *topic collapsing*, *i.e.*, topics are semantically collapsed together as illustrated in Table 1. We in this section investigate what causes the topic collapsing issue by analyzing a kind of recently proposed state-of-the-art neural topic models. Earlier neural topic models take the topic-word distribution matrix $\boldsymbol{\beta}$ directly as one parameter (Miao et al., 2016; Srivastava & Sutton, 2017). Recent state-of-the-art methods compute it with two parameters: $\boldsymbol{\beta} = \mathbf{W}^\top \mathbf{T}$ (Miao et al., 2017; Dieng et al., 2020; Zhao et al., 2021b). Here $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_V) \in \mathbb{R}^{D \times V}$ are the embeddings of all V words and $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_K) \in \mathbb{R}^{D \times K}$ are the embeddings of all K topics, all in the same D-dimensional semantic space. In this way, they can facilitate learning by initializing $\mathbf{W}$ with pre-trained embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).

However, the topic collapsing still exists although this initialization provides a good optimization starting point. To investigate the reason, we visualize the word and topic embeddings $\mathbf{W}$ and $\mathbf{T}$ learned by state-of-the-art ETM (Dieng et al., 2020), NSTM (Zhao et al., 2021b) and WeTe (Wang et al., 2022) through t-SNE (van der Maaten & Hinton, 2008). Figure 1 shows that the topic embeddings mostly collapse together in the semantic space of all models. This makes the topic-word distributions $\boldsymbol{\beta}_k$ very similar to each other, which leads to the topic collapsing. We argue the underlying reason is that these methods mainly minimize the reconstruction error between topic distribution $\boldsymbol{\theta}$ and word distribution of a document $\mathbf{x}$ to learn the word and topic embeddings. For example, ETM measures the reconstruction error with negative expected log-likelihood, NSTM uses optimal transport distance and WeTe uses conditional transport distance. However, it is easy for these methods to be trapped in bad local minima where they only reconstruct dominant words. Thus most topic embeddings approach to the embeddings of these words in the space, leading to topic collapsing.

3.3 WHAT IS AN EFFECTIVE CLUSTERING REGULARIZATION ON EMBEDDINGS?

In this section, we explore how to design an effective regularization on topic and word embeddings to address the topic collapsing issue. From the analysis in Sec. 3.2, we know the topic collapsing easily happens if only minimizing the reconstruction error. To address this issue, we intuitively propose

[1] Here Δ_K means a probability simplex $\Delta_K = \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^K \,|\, \sum_{k=1}^K \theta_k = 1 \right\}$.



(a) DKM (b) DKM + Entropy (c) ECR (ε=1.0) (d) ECR (ε=0.1) (e) ECR (ε=0.05) (f) Cluster size proportions.

Figure 2: t-SNE visualization (a-e) of word embeddings (●) and topic embeddings (▲) under 50 topics (K=50). (a): DKM cannot form effective clusters. (b): DKM + Entropy forms clusters but has a trivial solution that most clusters are empty. (c,d,e): Our ECR forms clusters with a small ε and avoids the trivial solution of empty clusters. (f): While most cluster size proportions are zero in DKM + Entropy, our ECR avoids this trivial solution by fulfilling cluster size constraints. Here all cluster sizes have equal proportions, $1/K$=0.02.

a clustering regularization on the topic and word embeddings beyond the reconstruction error. We require the regularization to consider topic embeddings as cluster centers and word embeddings as cluster samples. With an effective regularization, an ideal situation is that topic embeddings are away from each other and are the cluster centers of distinct word embedding clusters. So the topic-word distributions will be different from each other, and the discovered topics will contain distinct words, which alleviates the topic collapsing issue. However, it is non-trivial to achieve such an effective regularization on topic and word embeddings, and we explore the precise requirements as follows.

Supporting Joint Optimization As we regularize on neural topic models, we require the clustering regularization to support joint optimization on topic and word embeddings along with the neural topic modeling objective. Some studies (Sia et al., 2020) have applied classical clustering algorithms, *e.g.*, KMeans and GMM, to produce topics by clustering pre-trained word embeddings. We clarify that they are not topic models as they only produce topics and cannot learn doc-topic distributions as required (but we still compare them in experiments). We do not adopt these classical clustering algorithms and some other work (Song et al., 2013; Huang et al., 2014; Xie et al., 2016; Hsu & Lin, 2017; Yang et al., 2017) as our clustering regularization, because they cannot be jointly optimized along with the neural topic modeling objective.

Producing Sparse Soft-assignments We also require the clustering regularization to produce sparse soft-assignments. To support joint optimization, we propose to employ the state-of-the-art deep clustering method, Deep KMeans (DKM, Fard et al., 2020). The clustering objective of DKM is to minimize the total Euclidean distance between centers and samples weighted by soft-assignments, and we use it to formulate a clustering regularization on topic and word embeddings:

$$\min_{\mathbf{W},\mathbf{T},\mathbf{p}} \sum_{j=1}^{V}\sum_{k=1}^{K} \|\mathbf{w}_j - \mathbf{t}_k\|^2 p_{jk}, \quad \text{where} \quad p_{jk} = \frac{e^{-\|\mathbf{w}_j - \mathbf{t}_k\|^2/\tau}}{\sum_{k'=1}^{K} e^{-\|\mathbf{w}_j - \mathbf{t}_{k'}\|^2/\tau}}. \tag{1}$$

Here p_{jk} denotes the soft-assignment of word embedding $\mathbf{w}_j$ assigned to topic embedding $\mathbf{t}_k$ and is modeled as a parameterized softmax function of the Euclidean distance between $\mathbf{w}_j$ and all topic embeddings (τ is the temperature hyper-parameter). Unfortunately, Figure 2a shows that DKM cannot form effective clusters and still suffers from topic collapsing (quantitative results are in Sec. 4.3). We argue this is because their soft-assignments are not sparse enough to form effective clusters. Sparse soft-assignments indicate that each sample is mainly assigned to only one center and is rarely assigned to others. On the contrary, we see from Figure 2a that word embeddings (samples) are *not* close enough to their assigned topic embeddings (centers) and are *not* far away enough from other topic embeddings. Owing to this, we also require the clustering regularization to produce sparse soft-assignments.

Fulfilling Cluster Size Constraints We further require the clustering regularization to fulfill cluster size constraints. To make the soft-assignments sparse, we consider minimizing the entropy of soft-assignments. Specifically, we equip DKM with the entropy of soft-assignments $\sum_{j=1}^{V}\sum_{k=1}^{K} -p_{jk}\log p_{jk}$ and jointly minimizing it with the clustering objective of DKM (Eq. (1)).

Algorithm 1 Training algorithm for ECRTM.

Input: document collection $\mathbf{X}$, number of epochs n_{epoch};
Output: model parameters Θ, $\mathbf{W}$, $\mathbf{T}$;

for 1 to n_{epoch} **do**
  for every mini-batch $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N_B)})$ **do**
    // Sinkhorn's algorithm;
    $C_{jk} = \|\mathbf{w}_j - \mathbf{t}_k\|^2 \quad \forall\, j, k$;
    $\mathbf{M} = \exp(-\mathbf{C}/\epsilon)$;
    Initialize $\mathbf{b} \leftarrow \mathbb{1}_K$;
    while not converged and not reach max iterations **do**
      $\mathbf{a} \leftarrow \frac{1}{V}\frac{\mathbb{1}_V}{\mathbf{M}\mathbf{b}}$, $\mathbf{b} \leftarrow \frac{\mathbf{s}}{\mathbf{M}^\top \mathbf{a}}$;
    end while
    Compute $\boldsymbol{\pi}^*_\varepsilon = \text{diag}(\mathbf{a})\mathbf{M}\text{diag}(\mathbf{b})$;
    Compute $\mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}}\mathcal{L}_{\text{ECR}}$ (Eq. (7));
    Update Θ, $\mathbf{W}$, $\mathbf{T}$ with a gradient step;
  end for
end for

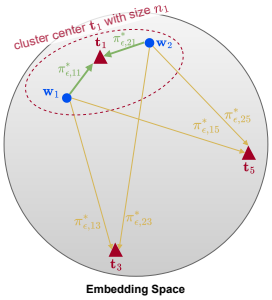


Figure 3: Illustration of ECR. ECR clusters word embeddings $\mathbf{w}_j$ (●) as samples and topic embeddings $\mathbf{t}_k$ (▲) as centers with soft-assignment $\pi^*_{\epsilon,jk}$. The cluster size of center $\mathbf{t}_k$ is constrained as n_k. Here ECR pushes $\mathbf{w}_1$ close to $\mathbf{t}_1$ and away from $\mathbf{t}_3$ and $\mathbf{t}_5$. It is similar for $\mathbf{w}_2$.

Figure 2b shows the soft-assignments indeed become sparser than DKM to form clusters, but unfortunately the clustering solution is trivial: most clusters are totally empty, which is quantitatively shown in Figure 2f. The topic embeddings (centers) of these empty clusters cannot learn to include different word embeddings. In consequence, this trivial clustering solution still leads to topic collapsing (quantitative results are in Sec. 4.3). To avoid such trivial solutions of empty clusters, we propose to preset constraints on the size of each cluster and further require the clustering regularization to fulfill the cluster size constraints.

3.4 Embedding Clustering Regularization

To meet the above requirements for solving topic collapsing, we propose a novel method, **Embedding Clustering Regularization** (**ECR**). Figure 3 shows the illustration of ECR, and Figure 2e shows its effectiveness.

Presetting Cluster Size Constraints We first preset the cluster size constraints to be fulfilled. Recall that we have K topic embeddings as centers and V word embeddings as samples. We require the cluster size of topic embedding $\mathbf{t}_k$ to be n_k, and $s_k = n_k/V$ represents the proportion of $\mathbf{t}_k$. We have $\mathbf{s} = (s_1, \ldots, s_K)^\top \in \Delta_K$ as the proportion vector of all clusters. Here we assume each cluster is the same size, $n_k = V/K$, so $\mathbf{s}$ is a uniform distribution $\mathbf{s} = (1/K, \ldots, 1/K)^\top$. This is because we often have no prior knowledge on cluster sizes, and it is reasonable to assume each topic includes the same amount of information without loss of generality. [2] Experiments also show this assumption works pretty well.

Embedding Clustering Regularization (ECR) To meet the above requirements of supporting joint optimization, producing sparse soft-assignments, and fulfilling cluster size constraints, we propose ECR which converts the clustering problem into a specifically defined entropic regularized optimal transport problem. To regularize topic and word embeddings, we define two discrete measures of them: $\gamma = \sum_{j=1}^{V} \frac{1}{V}\delta_{\mathbf{w}_j}$ and $\phi = \sum_{k=1}^{K} s_k \delta_{\mathbf{t}_k}$, where δ_x denotes the Dirac unit mass on x. We formulate the entropic regularized optimal transport problem between γ and ϕ as:

$$\underset{\boldsymbol{\pi} \in \mathbb{R}_+^{V\times K}}{\arg\min}\, \mathcal{L}_{\text{OT}_\varepsilon}(\gamma, \phi), \quad \mathcal{L}_{\text{OT}_\varepsilon}(\gamma, \phi) = \sum_{j=1}^{V}\sum_{k=1}^{K} \|\mathbf{w}_j - \mathbf{t}_k\|^2 \pi_{jk} + \sum_{j=1}^{V}\sum_{k=1}^{K} \varepsilon \pi_{jk}(\log(\pi_{jk}) - 1),$$
$$\text{s.t.} \quad \boldsymbol{\pi}\mathbb{1}_K = \frac{1}{V}\mathbb{1}_V \quad \text{and} \quad \boldsymbol{\pi}^\top \mathbb{1}_V = \mathbf{s}. \tag{2}$$

[2] Note that $\mathbf{s}$ can be determined by possible prior knowledge from experts, and we leave this as future work.

Here π_{jk} denotes the transport weight from $\mathbf{w}_j$ to $\mathbf{t}_k$, and $\boldsymbol{\pi}\in\mathbb{R}_+^{V\times K}$ is the transport plan restricted by two conditions where $\mathbb{1}_K$ ($\mathbb{1}_V$) means a K (D) dimensional column vector of ones. We measure the transport cost between $\mathbf{w}_j$ and $\mathbf{t}_k$ by Euclidean distance $\|\mathbf{w}_j - \mathbf{t}_k\|^2$. In Eq. (2), we formulate the first term as the original optimal transport problem and the second term with weight parameter ε as the entropic regularization that makes the optimal transport problem tractable (Canas & Rosasco, 2012; Genevay et al., 2019). We use Eq. (2) to find the optimal transport plan that minimizes the total transport cost from word embeddings to topic embeddings in the semantic space restricted by two conditions. To obtain the optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$, we employ Sinkhorn's algorithm (Sinkhorn, 1964; Cuturi, 2013), a much faster scheme particularly suited to the execution of GPU (Peyré et al., 2019):

$$\boldsymbol{\pi}_\varepsilon^* = \mathrm{sinkhorn}(\gamma, \phi, \varepsilon) \approx \underset{\boldsymbol{\pi}\in\mathbb{R}_+^{V\times K}}{\arg\min}\, \mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma, \phi). \tag{3}$$

Algorithm 1 shows the detail of Sinkhorn's algorithm. To meet the requirements of clustering regularization, we propose to take optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$ of Eq. (2) as the clustering soft-assignments in our ECR:

$$\mathcal{L}_{\mathrm{ECR}} = \sum_{j=1}^{V}\sum_{k=1}^{K} \|\mathbf{w}_j - \mathbf{t}_k\|^2 \pi_{\varepsilon,jk}^*. \tag{4}$$

ECR is an Effective Clustering Regularization As shown in Figures 2e and 2f, our ECR is an effective clustering regularization that meets the requirements in Sec. 3.2. First, we calculate $\boldsymbol{\pi}_\varepsilon^*$ through Sinkhorn's algorithm and make $\boldsymbol{\pi}_\varepsilon^*$ a differentiable variable with $\mathbf{W}$ and $\mathbf{T}$ as parameters (Salimans et al., 2018; Genevay et al., 2018). Thus the objective of ECR (Eq. (4)) can be jointly optimized with a neural topic modeling objective. Second, it has been proven that the solution $\boldsymbol{\pi}_\varepsilon^*$ converges to the optimal solution of the original optimal transport problem when $\varepsilon \to 0$, which leads to a sparse transport plan (Peyré et al., 2019). We illustrate this property in Figures 2c to 2e. So we can obtain sparse soft-assignments from the transport plan of a small ε to form effective clusters. Last, we restrict the transport plan by two conditions (Eq. (2)). The first condition $\boldsymbol{\pi}\mathbb{1}_K = \frac{1}{V}\mathbb{1}_V$ guarantees that each word embedding (cluster sample) is equally weighted as $\frac{1}{V}$. The second condition $\boldsymbol{\pi}^\top\mathbb{1}_V = \mathbf{s}$ ensures that the cluster size proportion of topic embedding $\mathbf{t}_k$ is s_k. These two conditions mean each topic embedding $\mathbf{t}_k$ is assigned with n_k word embeddings, such that $n_k \times \frac{1}{V} = s_k$ as shown in Figure 2f. This enables the soft-assignments to fulfill our preset cluster size constraints, and thus we can avoid trivial clustering solutions of empty clusters. To sum up, our ECR successfully meets the proposed requirements on an effective clustering regularization for alleviating the topic collapsing issue.

3.5 Neural Topic Modeling with Embedding Clustering Regularization

In this section, we propose a novel neural topic model, named **Embedding Clustering Regularization Topic Model** (**ECRTM)** that jointly optimizes the neural topic modeling objective and our proposed clustering regularization objective. Algorithm 1 is the training algorithm of ECRTM.

Inferring Topic Distributions We devise the prior and variational distribution following VAE Kingma & Welling (2014) to infer topic distributions. Since it is difficult to apply the reparameterization trick of VAE on the Dirichlet distribution as in LDA, we employ a logistic normal prior on $\boldsymbol{\theta}$ to avoid this issue (Srivastava & Sutton, 2017). In detail, we draw a latent variable $\mathbf{r}$, from a logistic normal distribution: $p(\mathbf{r}) = \mathcal{LN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the mean and diagonal covariance matrix. Then, we take the form of an encoder network which outputs parameters of the variational distribution, the mean vector $\boldsymbol{\mu} = f_{\boldsymbol{\mu}}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(f_{\boldsymbol{\Sigma}}(\mathbf{x}))$. So the variational distribution is $q_\Theta(\mathbf{r}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where Θ denotes the parameters of $f_{\boldsymbol{\mu}}$ and $f_{\boldsymbol{\Sigma}}$. By applying the reparameterization trick (Kingma & Welling, 2014), we sample $\mathbf{r} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. We obtain the topic distribution $\boldsymbol{\theta}$ with a softmax function as $\boldsymbol{\theta} = \mathrm{softmax}(\mathbf{r})$.

Reconstructing Documents We then reconstruct the input documents with topic-word distribution matrix $\boldsymbol{\beta} \in \mathbb{R}^{V\times K}$. Recall that $\boldsymbol{\beta}$ indicates the weights between topics and words. Previous methods commonly model $\boldsymbol{\beta}$ either as a single matrix parameter or as the product of the topic and word embeddings. Differently, our method uses ECR as a clustering regularization on the topic and

Table 2: Topic quality results concerning topic coherence (C_V) and diversity (TD) under 50 and 100 topics (K=50 and K=100). The best is highlighted in **bold**.

Model	20NG				IMDB			
	$K = 50$		$K = 100$		$K = 50$		$K = 100$	
	C_V	TD	C_V	TD	C_V	TD	C_V	TD
LDA	0.388±0.006	0.669±0.017	0.402±0.002	0.570±0.012	0.347±0.004	0.789±0.034	0.345±0.006	0.667±0.068
KM	0.251±0.008	0.204±0.017	0.294±0.013	0.317±0.017	0.213±0.008	0.219±0.011	0.244±0.003	0.302±0.006
WLDA	0.378±0.007	0.375±0.023	0.369±0.004	0.273±0.018	0.311±0.007	0.053±0.008	0.320±0.003	0.069±0.003
ETM	0.375±0.007	0.704±0.024	0.369±0.001	0.573±0.016	0.346±0.005	0.557±0.014	0.341±0.005	0.371±0.017
NSTM	0.395±0.005	0.427±0.017	0.391±0.003	0.473±0.025	0.334±0.003	0.175±0.010	0.340±0.003	0.255±0.012
WeTe	0.383±0.008	0.958±0.009	0.352±0.008	0.742±0.025	0.365±0.010	0.933±0.023	0.293±0.018	0.638±0.024
ECRTM	**0.431±0.013**	**0.964±0.009**	**0.404±0.004**	**0.897±0.025**	**0.395±0.005**	**0.979±0.020**	**0.373±0.006**	**0.887±0.019**

Model	Yahoo Answer				AG News			
	$K = 50$		$K = 100$		$K = 50$		$K = 100$	
	C_V	TD	C_V	TD	C_V	TD	C_V	TD
LDA	0.359±0.007	0.843±0.012	0.359±0.005	0.602±0.030	0.364±0.006	0.864±0.027	0.349±0.005	0.696±0.018
KM	0.271±0.018	0.242±0.008	0.297±0.006	0.345±0.008	0.241±0.015	0.264±0.026	0.289±0.007	0.395±0.016
WLDA	0.333±0.007	0.322±0.024	0.338±0.006	0.292±0.021	0.384±0.004	0.410±0.018	0.378±0.004	0.323±0.035
ETM	0.354±0.003	0.719±0.022	0.353±0.002	0.624±0.024	0.364±0.002	0.819±0.017	0.371±0.005	0.773±0.019
NSTM	0.390±0.009	0.658±0.011	**0.387±0.001**	0.659±0.005	0.411±0.011	0.873±0.019	**0.421±0.003**	0.832±0.010
WeTe	0.367±0.010	0.878±0.020	0.353±0.007	0.544±0.019	0.380±0.007	0.924±0.037	0.363±0.005	0.827±0.023
ECRTM	**0.409±0.005**	**0.971±0.021**	0.384±0.007	**0.946±0.017**	**0.466±0.012**	**0.961±0.009**	0.416±0.003	**0.981±0.012**

word embeddings, so $\boldsymbol{\beta}$ needs to reflect the learned geometric structures between the word and topic embeddings. We do not directly model $\boldsymbol{\beta}$ as the clustering soft-assignments $\boldsymbol{\pi}_\varepsilon^*$ because $\boldsymbol{\pi}_\varepsilon^*$ is very sparse as mentioned before (most values are close to zero), and it cannot reflect the weights between all topics and words. Experiments also show this way is ineffective. Hence we need a less sparse $\boldsymbol{\beta}$ for reconstruction. To this end, we propose to model $\boldsymbol{\beta}$ as the less sparse soft-assignments following DKM (Fard et al., 2020):

$$\beta_{jk} = \frac{e^{-\|\mathbf{w}_j - \mathbf{t}_k\|^2/\tau}}{\sum_{k'=1}^{K} e^{-\|\mathbf{w}_j - \mathbf{t}_{k'}\|^2/\tau}} \tag{5}$$

where τ is the temperature hyper-parameter. This formulation can reflect the learned geometric structures between topic and word embeddings and the weights between all topics and words for reconstruction. With the topic distribution $\boldsymbol{\theta}$ and the topic-word distribution matrix $\boldsymbol{\beta}$, we routinely sample the reconstructed document from a Multinomial distribution $\mathrm{Multi}(\mathrm{softmax}(\boldsymbol{\beta}\boldsymbol{\theta}))$.

Overall Objective Function We now conclude the topic modeling objective function following the ELBO of VAE. Given a mini-batch of N_B documents $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_B)}\}$, we formulate the topic modeling objective as

$$\mathcal{L}_{\text{TM}} = \frac{1}{N_B} \sum_{i=1}^{N_B} -\mathbf{x}^{(i)\top} \log(\mathrm{softmax}(\boldsymbol{\beta}\boldsymbol{\theta}^{(i)})) + \mathbf{KL}\left[q_\Theta(\mathbf{r}^{(i)}|\mathbf{x}^{(i)}) \| p(\mathbf{r}^{(i)})\right]. \tag{6}$$

The first term is the reconstruction error, and the second term is the KL divergence between the prior and variational distributions, which can be calculated with an analytic expression (Srivastava & Sutton, 2017). The overall objective function of ECRTM is a combination of $\mathcal{L}_{\text{TM}}$ (Eq. (6)) and $\mathcal{L}_{\text{ECR}}$ (Eq. (4)), defined as follows:

$$\min_{\Theta, \mathbf{W}, \mathbf{T}} \mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}} \mathcal{L}_{\text{ECR}} \tag{7}$$

where λ_{ECR} is a weight hyperparameter. This overall objective enables to cluster the embeddings of related words together with topic embeddings as centers and avoids the collapsing of topic embeddings. This allows our ECRTM to alleviate the topic collapsing issue and learn coherent and diverse topics together with high-quality doc-topic distributions at the same time.

4 Experiments

4.1 Experiment Settings

Datasets We adopt the following benchmark document datasets for experiments: (i) 20 News Groups (**20NG**, Lang, 1995) is one of the most popular datasets for evaluating topic models, in-

Table 3: Document clustering results of Purity and NMI under 50 and 100 topics ($K = 50$ and K=100). The best scores are in **bold**.

Model	20NG				IMDB			
	K=50		K=100		K=50		K=100	
	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI
LDA	0.314±0.029	0.302±0.016	0.303±0.028	0.284±0.016	0.614±0.006	0.041±0.008	0.605±0.011	0.039±0.009
WLDA	0.233±0.008	0.157±0.007	0.292±0.009	0.207±0.008	0.589±0.008	0.011±0.002	0.602±0.005	0.013±0.001
ETM	0.347±0.026	0.319±0.027	0.394±0.013	0.339±0.010	0.660±0.004	0.038±0.002	0.648±0.009	0.037±0.002
NSTM	0.354±0.032	0.356±0.015	0.383±0.021	0.363±0.015	0.658±0.008	0.040±0.002	0.659±0.010	0.039±0.003
WeTe	0.258±0.016	0.293±0.024	0.338±0.032	0.348±0.028	0.604±0.031	0.038±0.016	0.589±0.006	0.025±0.001
ECRTM	**0.560±0.037**	**0.524±0.025**	**0.551±0.009**	**0.493±0.005**	**0.690±0.009**	**0.056±0.007**	**0.694±0.006**	**0.049±0.002**

Model	Yahoo Answer				AG News			
	$K = 50$		$K = 100$		$K = 50$		$K = 100$	
	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI
LDA	0.288±0.017	0.144±0.013	0.297±0.019	0.148±0.007	0.640±0.028	0.193±0.017	0.654±0.012	0.194±0.006
WLDA	0.255±0.007	0.084±0.005	0.303±0.010	0.127±0.010	0.580±0.011	0.151±0.010	0.653±0.005	0.188±0.004
ETM	0.405±0.015	0.192±0.008	0.428±0.006	0.208±0.005	0.679±0.012	0.224±0.015	0.674±0.006	0.204±0.006
NSTM	0.395±0.022	0.241±0.009	0.405±0.013	0.242±0.011	0.719±0.041	0.324±0.022	0.764±0.018	0.359±0.008
WeTe	0.389±0.019	0.252±0.015	0.444±0.016	0.269±0.012	0.609±0.036	0.245±0.028	0.699±0.008	0.271±0.009
ECRTM	**0.540±0.008**	**0.295±0.005**	**0.550±0.004**	**0.304±0.005**	**0.802±0.009**	**0.367±0.013**	**0.812±0.020**	**0.428±0.016**

cluding news articles with 20 labels; (ii) **IMDB** (Maas et al., 2011) is the movie reviews containing two labels (positive and negative); (iii) **Yahoo Answer** (Zhang et al., 2015) is the question titles, contents, and the best answers from the Yahoo website with 10 labels. (iv) **AG News** (Zhang et al., 2015) contains news titles and descriptions, divided into 4 categories like Sports and Business. The preprocessing details are described in Appendix A.1.

Evaluation Metrics For sufficient evaluation of topic models (Hoyle et al., 2021), we evaluate the quality of topics and doc-topic distributions. For topic quality, we consider **topic coherence** and **topic diversity**. Topic coherence measures whether the top words of discovered topics are coherent (Newman et al., 2010; Wang & Blei, 2011). We employ the widely-used coherence metric, Coherence Value (C_V) (Röder et al., 2015). Besides, topic diversity measures the differences between discovered topics to verify if topic collapsing happens. We use Topic Diversity (TD, Dieng et al., 2020), which computes the proportion of unique words in the discovered topics. Additionally, we conduct **document clustering tasks** with doc-topic distributions to evaluate their quality following Zhao et al. (2021b). We use the most significant topics in the topic distributions of testing documents as their clustering assignments and adopt the commonly-used Purity and NMI (Manning et al., 2008) for evaluation. Note that our intention is not to achieve state-of-the-art document clustering results but compare the quality of doc-topic distributions.

Baseline models We consider the following state-of-the-art models for comparison: (i) **LDA** (Blei et al., 2003), one of the most widely-used probabilistic topic models; (ii) **KM** (Sia et al., 2020), directly clustering word embeddings to produce topics. It cannot be used for document clustering since it does not infer the doc-topic distributions. (iii) **WLDA** (Nan et al., 2019), a WAE-based topic model; (iv) **ETM** (Dieng et al., 2020), a neural topic model which models the topic-word distribution matrix with word and topic embeddings; (v) **NSTM** (Zhao et al., 2021b), using the optimal transport distance to measure reconstruction error. (vi) **WeTe** (Wang et al., 2022), following NSTM and using conditional transport distance as reconstruction error. More model settings are reported in Appendix A.2.

4.2 Main Results of Quality of Topics and Doc-topic Distributions

Table 2 shows the topic quality results regarding C_V and TD, and Table 3 summarizes the clustering results of Purity and NMI. We have the following observations: **(i) ECRTM effectively mitigates the topic collapsing issue and generates higher-quality topics.** In Table 2, the much lower TD scores show that baselines commonly generate repetitive topics and thus suffer from the topic collapsing. As aforementioned, these repetitive topics are less useless for downstream tasks and hamper the interpretability of topic models. However, our ECRTM achieves significantly higher TD across

Table 4: Ablation study. While DKM, DKM with Entropy, w/o ECR all have low TD scores, our ECRTM achieves better Purity and NMI scores and much higher TD scores. The low TD scores mean most topics are repetitive and less useful, making high C_V scores meaningless.

Model	20NG				Yahoo Answer			
	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD
DKM	0.545	0.482	0.451	0.585	0.507	0.282	0.403	0.631
DKM + Entropy	0.222	0.148	0.469	0.503	0.252	0.092	0.433	0.592
w/o ECR	0.530	0.447	0.476	0.577	0.498	0.262	0.435	0.608
ECRTM	0.560	0.524	0.431	0.964	0.540	0.295	0.409	0.971

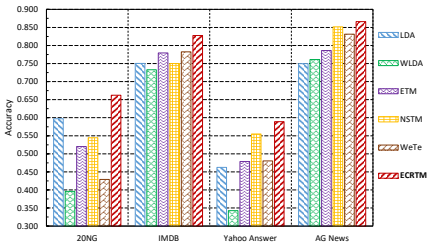


Figure 4: Text classification results with doc-topic distributions.

all datasets, and meanwhile the C_V scores of ECRTM are generally the best. Although the C_V of ECRTM is slightly lower than NSTM (0.384 v.s. 0.387) on Yahoo Answer, ECRTM achieves clearly better TD (0.946 v.s. 0.659). These demonstrate ECRTM has superior topic quality than state-of-the-art baselines. See some topic examples in Appendix A.3. (**ii**) **ECRTM surpasses baselines in inferring better doc-topic distributions as document representations.** As shown in Table 3, our ECRTM consistently outperforms the baseline models by a large margin in terms of Purity and NMI. For example, ECRTM reaches 0.560 and 0.524 for Purity and NMI on 20NG, while the runner-up only has 0.354 and 0.356. (**iii**) **ECRTM produces high-quality topics and doc-topic distributions at the same time.** From Tables 2 and 3, some baseline models perform well on document clustering but poorly on topic quality, such as NSTM on 20NG and ETM on Yahoo Answer. On the contrary, ECRTM not only achieves high-quality topics but also better doc-topic distributions as document representations.

4.3 Ablation Study

We conduct an ablation study and show the necessity of our proposed Embedding Clustering Regularization (ECR). Specifically, we remove the ECR of our ECRTM, denoted as w/o ECR. We also compare with the state-of-the-art deep clustering method, DKM (Fard et al., 2020) and DKM with minimizing entropy (see Sec. 3.3). Table 4 shows DKM and DKM + Entropy both suffer from topic collapsing as demonstrated by their much lower TD. Although w/o ECR has high C_V, its TD scores are greatly decreased compared to ECRTM (*e.g.*, from 0.964 to 0.577). This means topics of w/o ECR are highly repetitive and are less useful for downstream tasks, making the high C_V scores meaningless. Besides, ECRTM achieves the best document clustering performance with much higher Purity and NMI scores. These results demonstrate our ECR is effective and necessary to address the topic collapsing issue and improve the topic modeling performance.

4.4 Text Classification

To evaluate extrinsically, we further conduct text classification experiments as a downstream task. Specifically, we use the topic distributions learned by topic models as document representations and train SVMs to predict the class of each document. As reported in Figure 4, ECRTM significantly outperforms baseline models on all datasets (the p-values of significance tests are all less than 0.01). This shows that our method can be better utilized in this classification downstream task.

5 Conclusion

In this paper, we propose the novel and well-motivated Embedding Clustering Regularization Topic Model (ECRTM). ECRTM substantially improves the topic modeling performance and addresses the topic collapsing issue with the novel Embedding Clustering Regularization that effectively encourages topic embeddings to be the centers of distinct word embedding clusters. Extensive experiments demonstrate that ECRTM consistently achieves state-of-the-art performance in terms of producing high-quality topics and document representations. We hope our work can contribute to building reliable topic modeling applications.

REFERENCES

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

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

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Jordan L Boyd-Graber, Yuening Hu, David Mimno, et al. *Applications of topic models*, volume 11. now Publishers Incorporated, 2017.

Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27, 2019.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 25, 2012.

Dallas Card, Chenhao Tan, and Noah A Smith. Neural Models for Documents with Metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 2031–2040, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138:185–192, 2020.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Aude Genevay, Gabriel Dulac-Arnold, and Jean-Philippe Vert. Differentiable deep clustering with cluster size constraints. *CoRR*, abs/1910.09036, 2019. URL http://arxiv.org/abs/1910.09036.

Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. Kernel topic models. In *Artificial Intelligence and Statistics*, pp. 511–519, 2012.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

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

Chih-Chung Hsu and Chia-Wen Lin. Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data. *IEEE Transactions on Multimedia*, 20(2): 421–429, 2017.

Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. Deep embedding network for clustering. In *2014 22nd International conference on pattern recognition*, pp. 1532–1537. IEEE, 2014.

Viet Huynh, He Zhao, and Dinh Phung. Otlda: A geometry-aware optimal transport approach for topic modeling. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18573–18582. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d800149d2f947ad4d64f34668f8b20f6-Paper.pdf.

Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 567–576, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The International Conference on Learning Representations (ICLR)*, 2014.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 265–274. ACM, 2012. ISBN 1450311563.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172, 2013.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pp. 2121–2132, 2020.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pp. 1727–1736, 2016.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2410–2419. JMLR. org, 2017.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 880–889, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1092.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6345–6381, Florence, Italy, July 2019. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010. ISBN 1932432655.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *In Proceedings ofthe 31th International Conference on Machine Learning*, 2014.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408. ACM, 2015.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1105–1114. International World Wide Web Conferences Steering Committee, 2018.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1728–1736, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.135. URL `https://aclanthology.org/2020.emnlp-main.135`.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Chunfeng Song, Feng Liu, Yongzhen Huang, Liang Wang, and Tieniu Tan. Auto-encoder based data clustering. In *Proceedings, Part I, of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications-Volume 8258*, pp. 117–124, 2013.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BybtVK9lg`.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 448–456. ACM, 2011. ISBN 1450308139.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zhang, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=IYMuTbGzjFU`.

Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pp. 697–702. IEEE, 2007.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. Learning Multilingual Topics with Neural Variational Inference. In *International Conference on Natural Language Processing and Chinese Computing*, 2020.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.

Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 749–757. SIAM, 2013.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pp. 3861–3870. PMLR, 2017.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657, 2015.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. *arXiv preprint arXiv:2204.09874*, 2022.

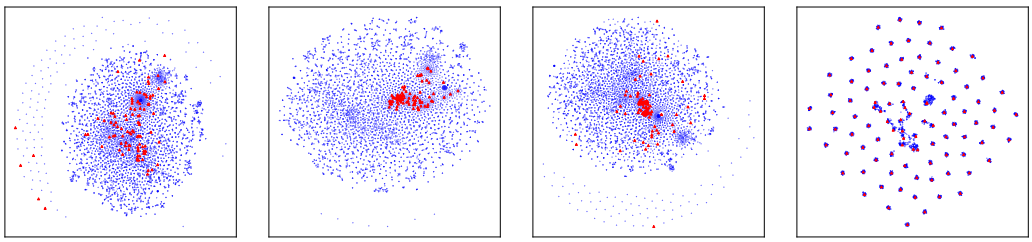
He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*, 2021a.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray L. Buntine. Neural topic model via optimal transport. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL `https://openreview.net/forum?id=Oos98K9Lv-k`.

A APPENDIX

Table 5: Top 10 related words of discovered topics from IMDB. Repetitive words are underlined.

ETM	like better good especially end look much done way just
	like just one way made much times really even feel
	one like around sort looking kind good main look just
NSTM	just show even come time one good really going know
	just even really something come going like actually things get
	just one even something come way really like always good
WeTe	hitchcocks couldve wouldve shouldve familys wifes anyones everyones itll hollywood
	hitchcocks anyones wouldve everyones wifes couldve familys itll shouldve hollywood
	hitchcocks couldve wouldve shouldve wifes familys anyones everyones itll hollywood
ECRTM	jackie martial chan kung arts kong hong stunts bruce fight
	nominated nancy academy award awards oscar oscars jake nomination dracula
	vampires vampire freddy zombies zombie nightmare serial halloween killer slasher



(a) ETM (b) NSTM (c) WeTe (d) **ECRTM**

Figure 5: t-SNE (van der Maaten & Hinton, 2008) visualization of word embeddings (●) and topic embeddings (▲) under 100 topics. Topic embeddings commonly collapse together in state-of-the-art models (ETM (Dieng et al., 2020), NSTM (Zhao et al., 2021b), and WeTe (Wang et al., 2022)). In contrast, ECRTM can avoid the collapsing of topic embeddings by effectively forming word embedding clusters.

A.1 DATASET

We follow the dataset preprocessing steps of Card et al. (2018): (1) documents are tokenized and converted to lower case; (2) remove punctuation; (3) remove tokens that include numbers; (4) remove tokens less than 3 characters; (5) remove stopwords according to the snowball sampler [3];

A.2 IMPLEMENTATION DETAILS

For pre-trained word embeddings, we employ 200-dimensional GloVe (Pennington et al., 2014) [4]. For the Sinkhorn's algorithm of ECRTM, we set the maximum number of iterations as 1,000, the stop tolerance 0.005 and ε 0.05 following Cuturi (2013). For ECRTM, the prior distribution is specified with Laplace approximation (Hennig et al., 2012) to approximate a symmetric Dirichlet prior as $\mu_{0,k} = 0$ and $\Sigma_{0,kk} = (K-1)/(\alpha K)$ with hyperparameter α. The α in the prior distribution is

[3] http://snowball.tartarus.org/algorithms/english/stop.txt
[4] https://nlp.stanford.edu/projects/glove/

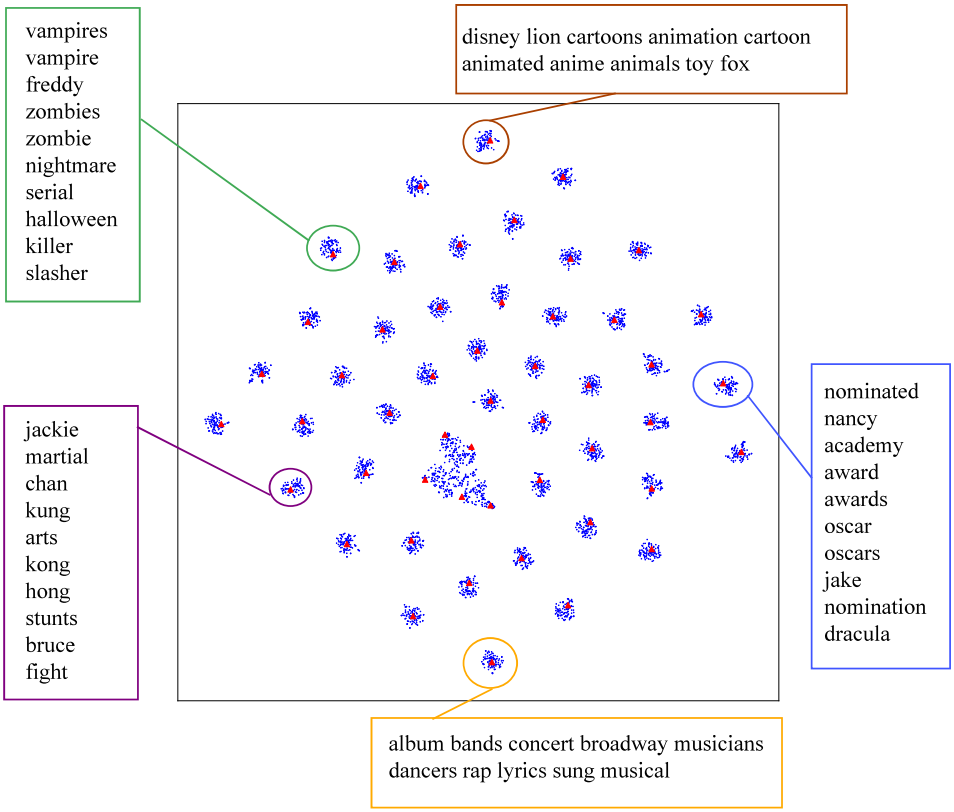


Figure 6: Visualization of discovered topics in the embedding space.

1.0 following Card et al. (2018) and τ is 0.2. Besides, λ_{ECR} is 250, 100, 70 and 5 for 20NG, IMDB, Yahoo Answer and AG News respectively as it is related to the average length of datasets. Our encoder network is the same as Srivastava & Sutton (2017), a MLP that has two linear layers with softplus activation function, concatenated with two single layers each for the mean and covariance matrix. We use Adam optimizer (Kingma & Ba, 2014) to optimize the model parameters.

A.3 Examples of Discovered Topics

Table 5 shows randomly selected examples of discovered topics by different models from IMDB. ETM and NSTM both have highly uninformative and similar topics including common words like "just", "like" or "something". WeTe produces some exactly the same topics with words "couldve", "wouldve", and "shouldve". These topics are uninformative and redundant, which are less useful for downstream task and damage the interpretability of topic models. In contrast, the topics discovered by ECRTM are more distinct instead of repeating each other. Besides, they are relatively more coherent, such as the first topic with relevant words like "jackie", "chan" and "stunts". These examples show our ECRTM generates higher-quality topics.

A.4 Visualization of Embeddings

We visualize the learned topic and word embeddings with t-SNE (van der Maaten & Hinton, 2008) under 100 topics. Figure 5 shows while the topic embedding mostly collapse together in the state-of-the-art baselines, our ECRTM effectively form distinct word embedding clusters with topic embeddings as centers.

We further annotate the embedding space with the top related words of discovered topics as shown in Figure 6. We see that each word embedding cluster represents a diverse and coherent topic. This verifies that our ECRTM effectively clusters the embeddings of coherent words by jointly optimizing the neural topic modeling objective with the embedding clustering regularization objective.

A.5 RUNTIME ANALYSIS

All models are trained using NVIDIA GPU and our implementation is based on PyTorch. It takes about less than 0.5 GPU hours to train our model on each dataset.