Effective Neural Topic Modeling with Embedding Clustering Regularization

Xiaobao Wu [1] **Xinshuai Dong** [2] **Thong Nguyen** [3] **Anh Tuan Luu** [1]

Abstract

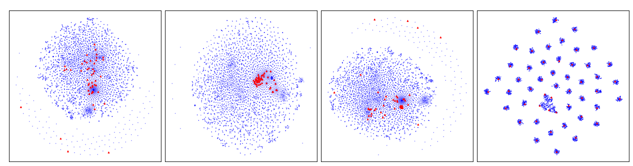
Topic models have been prevalent for decades with various applications. However, existing topic models commonly suffer from the notorious topic collapsing: discovered topics semantically collapse towards each other, leading to highly repetitive topics, insufficient topic discovery, and damaged model interpretability. In this paper, we propose a new neural topic model, Embedding Clustering Regularization Topic Model (ECRTM). Besides the existing reconstruction error, we propose a novel Embedding Clustering Regularization (ECR), which forces each topic embedding to be the center of a separately aggregated word embedding cluster in the semantic space. This enables each produced topic to contain distinct word semantics, which alleviates topic collapsing. Regularized by ECR, our ECRTM generates diverse and coherent topics together with high-quality topic distributions of documents. Extensive experiments on benchmark datasets demonstrate that ECRTM effectively addresses the topic collapsing issue and consistently surpasses state-of-the-art baselines in terms of topic quality, topic distributions of documents, and downstream classification tasks.

1. Introduction

Topic models have achieved great success in document analysis via discovering latent semantics. They have facilitated 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: (1) conventional topic models with probabilistic graphical models (Blei et al., 2003) or matrix factorization (Kim et al., 2015; Shi et al., 2018) and (2) neural topic models (Miao et al., 2016; 2017; Srivastava & Sutton, 2017; Gupta et al., 2019).

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: Top related words of the discovered topics by NSTM (Zhao et al., 2021b) on IMDB. These topics semantically collapse towards each other with many uninformative and repetitive words. Repetitions are underlined.



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

Figure 1: t-SNE visualization of word embeddings (●) and topic embeddings (▲) under 50 topics. These show while the topic embeddings mostly collapse together in previous state-of-the-art models, our ECRTM successfully avoids the collapsing.

However, despite the current achievements, existing topic models commonly suffer from *topic collapsing*: the discovered topics tend to semantically collapse towards each other (Srivastava & Sutton, 2017), as exemplified in Table 1. We see these collapsed topics include many uninformative and repetitive words. This brings about several issues: (**1**) Topic collapsing results in highly repetitive topics, which are less useful for downstream applications (Wallach et al., 2009; Nan et al., 2019). (**2**) Topic collapsing incurs insufficient topic discovery. Many latent topics are undisclosed, making the topic discovery insufficient to understand documents (Dieng et al., 2020). (**3**) Topic collapsing damages the interpretability of topic models. It becomes difficult to infer the real underlying topics that a document contains (Huynh et al., 2020). In consequence, topic collapsing impedes the utilization and extension of topic mod-

[1]Nanyang Technological University [2]Carnegie Mellon University [3]National University of Singapore. Correspondence to: Xiaobao Wu <xiaobao002@e.ntu.edu.sg>, Anh Tuan Luu <anhtuan.luu@ntu.edu.sg>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

els; therefore it is crucial to overcome this challenge for building effective topic models.

To address the topic collapsing issue and achieve effective topic modeling, we in this paper propose a novel neural topic model, **Embedding Clustering Regularization Topic Model** (**ECRTM**). First, we illustrate the reason for topic collapsing: Figures 1a to 1c show topic embeddings mostly collapse together in the semantic space of previous state-of-the-art methods. This makes discovered topics contain similar word semantics and thus results in the topic collapsing. Then to avoid the collapsing of topic embeddings, we propose the novel **Embedding Clustering Regularization** (**ECR**) besides the reconstruction error of previous work. ECR regularizes topic embeddings as cluster centers and word embeddings as cluster samples. For effective regularization, ECR models the clustering soft-assignments between them by solving a specifically defined optimal transport problem on them. As such, ECR forces each topic embedding to be the center of a separately aggregated word embedding cluster. Instead of collapsing together, this makes topic embeddings away from each other and cover different semantics of word embeddings. Thus our ECR enables each produced topic to contain distinct word semantics, which alleviates topic collapsing. Regularized by ECR, our ECRTM achieves effective topic modeling performance by producing diverse and coherent topics together with high-quality topic distributions of documents. Figure 1d shows the effectiveness of ECRTM. We conclude the main contributions of our paper as follows [1]:

- We propose a novel embedding clustering regularization that avoids the collapsing of topic embeddings by forcing each topic embedding to be the center of a separately aggregated word embedding cluster, which effectively mitigates topic collapsing.
- We further propose a new neural topic model that jointly optimizes the topic modeling objective and the embedding clustering regularization objective. Our model can produce diverse and coherent topics with high-quality topic distributions of documents at the same time.
- We conduct extensive experiments on benchmark datasets and demonstrate that our model effectively addresses 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; Yao et al., 2014; Das et al., 2015) mostly employ probabilistic graphical models to model the generation 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). Some studies use matrix factorization to model topics (Yan et al., 2013b; Kim et al., 2015; Shi et al., 2018). Many various scenarios have been developed, like short texts (Yan et al., 2013a; Wu & Li, 2019), multilingual (Mimno et al., 2009), and dynamic topic models (Blei & Lafferty, 2006). These methods commonly need model-specific derivations for different modeling assumptions.

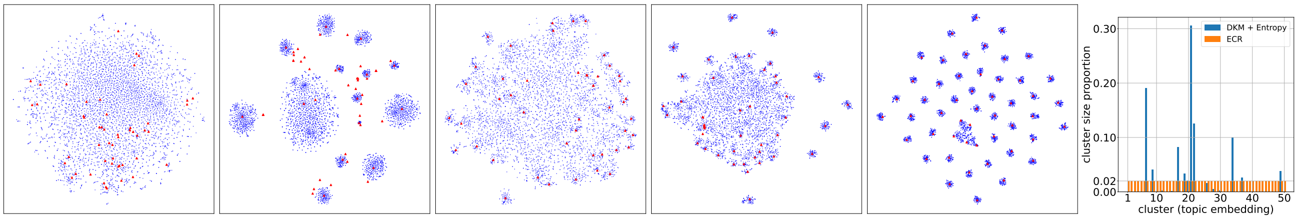
Neural Topic Models Due to the success of Variational AutoEncoder (VAE, Kingma & Welling, 2014; Rezende et al., 2014), several neural topic models have been proposed (Miao et al., 2016; Srivastava & Sutton, 2017; Dieng et al., 2019; Meng et al., 2020; Nguyen & Luu, 2021; Wu et al., 2021; 2022; 2023). Different from conventional ones, neural topic models can directly apply gradient backpropagation, which enhances flexibility and scalability. Alternatively, 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 state-of-the-art NSTM (Zhao et al., 2021b) and WeTe (Wang et al., 2022) measure the reconstruction error with optimal (conditional) transport distance. However, they still suffer from topic collapsing (see Sec. 4.2). Different from these, our proposed model aims to address the topic collapsing issue and achieve effective neural topic modeling. Besides the reconstruction error of these previous work, we propose a novel embedding clustering regularization that avoids the collapsing of topic embeddings by forcing each topic embedding to be the center of a separately aggregated word embedding cluster. Then our model learns topics under this effective regularization and particularly addresses the topic collapsing issue.

3. Methodology

3.1. Problem Setting and Notations

We recall the problem setting 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 latent topics from this document collection. The k-th 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 of all topics. The topic distribution of a document (doc-topic distribution) refers to what topics it contains, denoted as $\boldsymbol{\theta} \in \Delta_K$. Here Δ_K denotes a probability simplex $\Delta_K = \Big\{ \boldsymbol{\theta} \in \mathbb{R}_+^K \,|\, \sum_{k=1}^K \theta_k = 1 \Big\}$.

[1] Our code is available at https://github.com/bobxwu/ECRTM



(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 separately aggregated clusters. (b): DKM + Entropy forms clusters but has a trivial solution that clusters of most topic embeddings are empty. (c,d,e): Our ECR forms clusters and also avoids the trivial solution of empty clusters with a small ε. (f): This quantitatively shows that while most cluster size proportions are zero in DKM + Entropy, our ECR successfully avoids this trivial solution with fulfilled preset cluster size constraints. Here we preset all cluster sizes as equal, so cluster size proportions all are $1/K$=0.02 (See Sec. 3.4).

3.2. What Causes Topic Collapsing?

Despite the current achievements, most topic models suffer from *topic collapsing*: topics semantically collapse towards each other (see Table 1). We illustrate what causes topic collapsing by analyzing a kind of recently proposed state-of-the-art neural topic models (Dieng et al., 2020; Zhao et al., 2021b). These models compute the topic-word distribution matrix with two parameters: $\boldsymbol{\beta} = \mathbf{W}^{\top}\mathbf{T}$. Here $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_V) \in \mathbb{R}^{D \times V}$ are the embeddings of V words, and $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_K) \in \mathbb{R}^{D \times K}$ are the embeddings of K topics, all in the same D-dimensional semantic space. They can facilitate learning by initializing $\mathbf{W}$ with pre-trained embeddings like GloVe (Pennington et al., 2014).

However, topic collapsing commonly happens in these state-of-the-art models. We believe the reason lies in that their reconstruction error minimization incurs the collapsing of topic embeddings. Specifically, these models learn topic and word embeddings by minimizing the reconstruction error between topic distribution $\boldsymbol{\theta}$ and word distribution $\mathbf{x}$ of a document. For example, to measure reconstruction error, ETM (Dieng et al., 2020) uses traditional expected log-likelihood, and recent NSTM (Zhao et al., 2021b) and WeTe (Wang et al., 2022) use optimal (conditional) transport distance. In fact, words in a document collection commonly are long-tail distributed following Zipf's law (Reed, 2001; Piantadosi, 2014)—roughly speaking, few words are of high frequency and most are of low frequency. Therefore the reconstruction is biased as it mainly reconstructs high-frequency words regardless of the reconstruction error measurements. Since topic and word embeddings are learned by minimizing reconstruction error, this biased reconstruction pushes most topic embeddings close to the embeddings of some high-frequency words in the semantic space. As a result, topic embeddings collapse together in these state-of-the-art methods as shown in Figure 1. The topic-word distributions become similar to each other, leading to topic collapsing. We empirically demonstrate this argument by removing high-frequency words (See experiments in Sec. 4.6).

3.3. How to Design An Effective Regularization on Embeddings?

In this section, we discuss how to design an effective regularization on embeddings for the topic collapsing issue.

Our analysis in Sec. 3.2 indicates topic collapsing happens because the reconstruction error minimization incurs the collapsing of topic embeddings in existing work. To address this issue, we propose to design a clustering regularization on embeddings in addition to the reconstruction error of existing work. We consider topic embeddings as cluster centers and word embeddings as cluster samples; then we require the regularization to force each topic embedding to be the center of a separately aggregated word embedding cluster. As such, instead of collapsing together, topic embeddings are away from each other and cover different semantics of word embeddings in the space. This will make each discovered topic contain distinct word semantics and thus alleviate topic collapsing. **However, it is *non-trivial* to design such an effective regularization.** We explore the 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 a neural topic modeling objective. Some studies (Sia et al., 2020) apply classic 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

Algorithm 1 Training algorithm for ECRTM.

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

for 1 to n_{epoch} **do**
  for each mini-batch $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$ from $\mathbf{X}$ **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 $\boldsymbol{\pi}^*_\varepsilon \leftarrow \text{diag}(\mathbf{a})\mathbf{M}\text{diag}(\mathbf{b})$;
    Compute $\mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}}\mathcal{L}_{\text{ECR}}$ (Eq. (6));
    Update Θ, $\mathbf{W}$, $\mathbf{T}$ with a gradient step;
  end for
end for

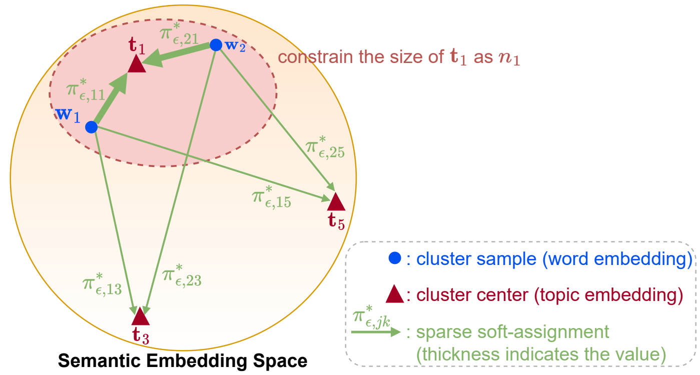


Figure 3: Illustration of ECR. ECR clusters word embeddings $\mathbf{w}_j$ (●) as samples and topic embeddings $\mathbf{t}_k$ (▲) as centers with soft-assignment $\pi^*_{\epsilon,jk}$. The cluster size of center $\mathbf{t}_k$ is constrained as n_k. Here ECR pushes $\mathbf{w}_1$ and $\mathbf{w}_2$ close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

compare them in experiments). We do not adopt these classic 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 we cannot jointly optimize them along with a neural topic modeling objective.

Producing Sparse Soft-assignments We also require the clustering regularization to produce sparse soft-assignments. Even supporting joint optimization, existing clustering methods may still lead to topic collapsing. For example, we propose to use the state-of-the-art deep clustering method, Deep KMeans (DKM, Fard et al., 2020) that supports joint optimization. **Note we are the first to use DKM in topic modeling.** Its clustering objective is to minimize the total Euclidean distance between centers and samples weighted by soft-assignments. We use DKM as a clustering regularization on topic and word embeddings:

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

Here p_{jk} denotes the clustering soft-assignment of word embedding $\mathbf{w}_j$ assigned to topic embedding $\mathbf{t}_k$, which is modeled as a softmax function of the Euclidean distance between $\mathbf{w}_j$ and all topic embeddings (τ is a temperature hyperparameter). Unfortunately, DKM still incurs topic collapsing (See quantitative results in Sec. 4.3). We see from Figure 2a that DKM cannot form separately aggregated clusters, so the topic embeddings (centers) cannot be separated but collapse together. To solve this issue, we require the clustering regularization to produce sparse soft-assignments—each word embedding is mainly assigned to only one topic embedding and rarely to others, which pushes each word embedding only close to one topic embedding and away from all others in the semantic space. This way can form separately aggregated word embedding clusters with topic embeddings as centers, which encourages topic embeddings to be away from each other. **Note that we do *not* directly model latent topics with these sparse soft-assignments (See Sec. 3.5).**

Fulfilling Preset Cluster Size Constraints We further require the clustering regularization to fulfill preset cluster size constraints. Only producing sparse soft-assignments may still result in topic collapsing. To make the soft-assignments sparse, we propose DKM+Entropy that jointly minimizes the entropy of soft-assignments, $\sum_{j=1}^{V}\sum_{k=1}^{K} -p_{jk}\log p_{jk}$, with the clustering objective of DKM (Eq. (1)). However, this way still leads to topic collapsing (See quantitative results in Sec. 4.3). Figure 2b shows DKM+Entropy indeed forms separately aggregated clusters for some topic embeddings, but unfortunately the clustering solution is trivial—the clusters of most topic embeddings are empty, as quantitatively shown in Figure 2f. As a result, the topic embeddings of these empty clusters cannot be separated to cover distinct semantics but only collapse to others in the space. To avoid such trivial solutions of empty clusters, we propose to preset constraints on the size of each cluster (must not be empty) and require the clustering regularization to fulfill these constraints.

3.4. Embedding Clustering Regularization

To meet the above necessary requirements, we in this section introduce a novel method, **Embedding Clustering Regularization** (**ECR**). Figure 3 illustrates ECR, and Figure 2e shows its effectiveness.

Presetting Cluster Size Constraints We first preset the cluster size constraints to be fulfilled to avoid trivial solutions of empty clusters. Recall that we have K topic embeddings as centers and V word embeddings as samples. We denote the cluster size of topic embedding $\mathbf{t}_k$ as n_k and the cluster size proportion as $s_k = n_k/V$. We have $\mathbf{s}=(s_1,\ldots,s_K)^\top\in\Delta_K$ as the vector of all cluster size proportions. Unfortunately, we usually lack prior knowledge about the cluster sizes of topic embeddings. Previous studies (Wallach et al., 2009) find that a symmetric Dirichlet prior over topic-word distributions achieves better performance in LDA. Inspired by this, we set all cluster sizes as uniform: n_k=V/K and $\mathbf{s}$=$(1/K,\ldots,1/K)^\top$. This setting can avoid the trivial solutions of empty clusters, and experiments show it works well across datasets (see Sec. 4.2). Note that $\mathbf{s}$ can be different values determined by prior knowledge from experts, and we leave this as future work.

Embedding Clustering Regularization (ECR) To meet the above requirements, we propose ECR that models clustering soft-assignments with the transport plan of a specifically defined optimal transport problem. Specifically, we define two discrete measures of topic ($\mathbf{t}_k$) and word embeddings ($\mathbf{w}_j$): γ=$\sum_{j=1}^{V}\frac{1}{V}\delta_{\mathbf{w}_j}$ and ϕ=$\sum_{k=1}^{K}s_k\delta_{\mathbf{t}_k}$, where δ_x denotes the Dirac unit mass on x. We formulate the entropic regularized optimal transport between γ and ϕ as

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

Here the first term of $\mathcal{L}_{\mathrm{OT}_\varepsilon}$ is the original optimal transport problem, and the second term with hyperparameter ε is the entropic regularization to make this problem tractable (Canas & Rosasco, 2012). Eq. (2) is to find the optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$ that minimizes the total cost of transporting weight from word embeddings to topic embeddings. We measure the transport cost between word embedding $\mathbf{w}_j$ and topic embedding $\mathbf{t}_k$ by Euclidean distance: $C_{jk}=\|\mathbf{w}_j-\mathbf{t}_k\|^2$, and the transport cost matrix is $\mathbf{C}\in\mathbb{R}^{V\times K}$. The two conditions in Eq. (2) restrict the weight of each word embedding $\mathbf{w}_j$ as $\frac{1}{V}$ and each topic embedding $\mathbf{t}_k$ as s_k, where $\mathbb{1}_K$ ($\mathbb{1}_V$) is a K (V) dimensional column vector of ones. π_{jk} denotes the transport weight from $\mathbf{w}_j$ to $\mathbf{t}_k$; $\boldsymbol{\pi}$$\in$$\mathbb{R}_+^{V\times K}$ is the transport plan that includes the transport weight of each word embedding to fulfill the weight of each topic embedding.

To meet the above requirements, we model clustering soft-assignments with the optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$, *i.e.*, the soft-assignment of $\mathbf{w}_j$ to $\mathbf{t}_k$ is the transport weight between them, $\pi_{\varepsilon,jk}^*$. Then we formulate our ECR objective by minimizing the total distance between word and topic embeddings weighted by these soft-assignments:

$$\begin{aligned}&\mathcal{L}_{\mathrm{ECR}}=\sum_{j=1}^{V}\sum_{k=1}^{K}\|\mathbf{w}_j-\mathbf{t}_k\|^2\pi_{\varepsilon,jk}^*,\\ &\text{where }\boldsymbol{\pi}_\varepsilon^*=\mathrm{sinkhorn}(\gamma,\phi,\varepsilon)\approx\underset{\boldsymbol{\pi}\in\mathbb{R}_+^{V\times K}}{\arg\min}\,\mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma,\phi).\end{aligned}\tag{3}$$

To solve this specifically defined optimal transport problem, here we compute $\boldsymbol{\pi}_\varepsilon^*$ through Sinkhorn's algorithm (Sinkhorn, 1964; Cuturi, 2013), a fast iterative scheme suited to the execution of GPU (Peyré et al., 2019). See Algorithm 1 for detailed algorithm steps. By doing so, $\boldsymbol{\pi}_\varepsilon^*$ is a differentiable variable parameterized by transport cost matrix $\mathbf{C}$ (Salimans et al., 2018; Genevay et al., 2018). Due to this, minimizing $\pi_{\varepsilon,jk}^*$ increases transport cost C_{jk}, *i.e.*, the distance between $\mathbf{w}_j$ and $\mathbf{t}_k$; otherwise maximizing it decreases the distance (Genevay et al., 2019). Thus we can exactly model $\boldsymbol{\pi}_\varepsilon^*$ as differentiable clustering soft-assignments between topic and word embeddings.

ECR is an Effective Regularization on Embeddings First, ECR supports joint optimization since $\boldsymbol{\pi}_\varepsilon^*$ is differentiable as aforementioned. Second, ECR produces sparse soft-assignments. It is proven that $\boldsymbol{\pi}_\varepsilon^*$ 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). Hence ECR (Eq. (3)) produces sparse soft-assignments under a small ε. With sparse soft-assignments, ECR pushes each word embedding only close to one topic embedding and away from all others, which forms separately aggregated clusters. We illustrate this property in Figures 2c to 2e. Last, ECR fulfills preset cluster size constraints. In Eq. (2), the transport plan is restricted by two conditions indicating the weight of each word embedding $\mathbf{w}_j$ is $\frac{1}{V}$ and each topic embedding $\mathbf{t}_k$ is s_k. These ensure the sparse optimal transport plan $\boldsymbol{\pi}_\varepsilon^*$ needs to transport n_k word embeddings to topic embedding $\mathbf{t}_k$ to balance the weight, such that $n_k\times\frac{1}{V}=s_k$. Accordingly, ECR fulfills the preset cluster size constraints with $\boldsymbol{\pi}_\varepsilon^*$ as clustering soft-assignments. This avoids trivial clustering solutions of empty clusters as shown in Figure 2f.

To sum up, our ECR effectively forces each topic embedding to be the center of a separately aggregated word embedding cluster. This makes topic embeddings away from each other and cover different semantics of word embeddings, which alleviates topic collapsing.

3.5. Embedding Clustering Regularization Topic Model

In this section, we propose a novel topic model, **Embedding Clustering Regularization Topic Model (ECRTM)** that jointly optimizes the topic modeling objective and the ECR objective. Algorithm 1 shows its training algorithm.

Inferring Doc-Topic Distributions We devise the prior and variational distribution following VAE (Kingma & Welling, 2014) to infer doc-topic distributions. 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 (Srivastava & Sutton, 2017). Then we use an encoder network that 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 doc-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 all topics and words. Previous methods commonly model $\boldsymbol{\beta}$ as the product of topic and word embeddings (Sec. 3.2). Differently, our model uses the proposed ECR as a clustering regularization on topic and word embeddings, so $\boldsymbol{\beta}$ also needs to reflect the learned clustering assignments between them. **We do *not* directly model $\boldsymbol{\beta}$ with the soft-assignments $\boldsymbol{\pi}^*_\varepsilon$ of our ECR.** This is because it makes one word only belong to one topic as $\boldsymbol{\pi}^*_\varepsilon$ is very sparse (most values are close to zero as aforementioned); but in reality, one word should be able to belong to different topics (Blei et al., 2003). To this end, we propose to model $\boldsymbol{\beta}$ as

$$\beta_{jk} = \frac{e^{-\|\mathbf{w}_j - \mathbf{t}_k\|^2/\tau}}{\sum_{k'=1}^{K} e^{-\|\mathbf{w}_j - \mathbf{t}_{k'}\|^2/\tau}} \tag{4}$$

where τ is a temperature hyperparameter. This formulation is similar to the less sparse soft-assignments of DKM (Fard et al., 2020). It is less sparse and can reflect the learned clustering assignments between topic and word embeddings. Thus one word can belong to different topics, which agrees with reality. With the doc-topic distribution $\boldsymbol{\theta}$ and the topic-word distribution matrix $\boldsymbol{\beta}$, we routinely sample the reconstructed document from a Multinomial distribution $\mathrm{Multi}(\mathrm{softmax}(\boldsymbol{\beta\theta}))$.

Overall Objective Function of ECRTM Given a batch of N documents $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$, we write the topic modeling objective function following VAE as

$$\mathcal{L}_{\mathrm{TM}} = \frac{1}{N} \sum_{i=1}^{N} -(\mathbf{x}^{(i)})^\top \log(\mathrm{softmax}(\boldsymbol{\beta\theta}^{(i)})) + \mathrm{KL}\left[q_\Theta(\mathbf{r}^{(i)}|\mathbf{x}^{(i)}) \| p(\mathbf{r}^{(i)})\right]. \tag{5}$$

The first term is the reconstruction error, and the second term is the KL divergence between the prior and variational distribution. ECRTM learns topics regularized by our ECR. We define the overall objective function of ECRTM as a combination of $\mathcal{L}_{\mathrm{TM}}$ (Eq. (5)) and $\mathcal{L}_{\mathrm{ECR}}$ (Eq. (3)):

$$\min_{\Theta, \mathbf{W}, \mathbf{T}} \mathcal{L}_{\mathrm{TM}} + \lambda_{\mathrm{ECR}}\, \mathcal{L}_{\mathrm{ECR}} \tag{6}$$

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

4. Experiment

4.1. Experiment Setup

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, including 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, such as Society, Culture, and Family & Relationships; (iv) **AG News** (Zhang et al., 2015) contains news titles and descriptions, divided into 4 categories like Sports and Business. Note that Yahoo Answer and AG News belong to short texts. See Appendix A for pre-processing details.

Evaluation Metrics Following previous mainstream work, we evaluate topic models concerning **topic quality and doc-topic distribution quality**. Topic quality includes: (i) **Topic Coherence** measures the coherence between the top words of discovered topics (Newman et al., 2010; Wang & Blei, 2011). We employ the widely-used metric, Coherence Value (C_V) which has been empirically shown to outperform the traditional metrics, NPMI, UCI, and UMass (Röder et al., 2015). We also exemplify this in Appendix E. We use the public Wikipedia article collection [2] as the external reference corpus. This removes the bias of using relatively small datasets (*e.g.*, training sets) as the reference corpus, so we can reach fair comparisons and good reproducibility. (ii) **Topic Diversity** measures the differences between discovered topics to verify if topic collapsing happens. We use the Topic Diversity metric (TD, Dieng et al., 2020) to evaluate this performance, which computes the proportion of unique words in the discovered topics. We select the top 15 words of discovered

[2] https://github.com/dice-group/Palmetto

Model	20NG				IMDB				Yahoo Answer				AG News			
	K=50		K=100		K=50		K=100		K=50		K=100		K=50		K=100	
	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD
LDA	‡0.385	‡0.655	‡0.387	‡0.622	‡0.347	‡0.788	‡0.342	‡0.691	‡0.359	‡0.843	‡0.359	‡0.602	‡0.364	‡0.864	‡0.349	‡0.696
KM	‡0.251	‡0.204	‡0.294	‡0.317	‡0.213	‡0.219	‡0.244	‡0.302	‡0.271	‡0.242	‡0.297	‡0.345	‡0.241	‡0.264	‡0.289	‡0.395
WLDA	‡0.378	‡0.375	‡0.369	‡0.273	‡0.311	‡0.053	‡0.320	‡0.069	‡0.333	‡0.322	‡0.338	‡0.292	‡0.384	‡0.410	‡0.378	‡0.323
DVAE	‡0.331	‡0.598	‡0.372	‡0.658	‡0.294	‡0.050	‡0.290	‡0.229	‡0.338	‡0.674	‡0.376	‡0.589	‡0.419	‡0.347	‡0.298	‡0.113
ETM	‡0.375	‡0.704	‡0.369	‡0.573	‡0.346	‡0.557	‡0.341	‡0.371	‡0.354	‡0.719	‡0.353	‡0.624	‡0.364	‡0.819	‡0.371	‡0.773
HyperMiner	‡0.371	‡0.613	‡0.368	‡0.446	‡0.347	‡0.485	‡0.343	‡0.258	‡0.344	‡0.507	‡0.346	‡0.444	‡0.359	‡0.521	‡0.360	‡0.343
NSTM	‡0.395	‡0.427	‡0.391	‡0.473	‡0.334	‡0.175	‡0.340	‡0.255	‡0.390	‡0.658	0.387	‡0.659	‡0.411	‡0.873	**0.421**	‡0.832
WeTe	‡0.383	‡0.949	‡0.352	‡0.742	‡0.368	‡0.931	‡0.293	‡0.638	‡0.367	‡0.878	‡0.353	‡0.544	‡0.383	‡0.945	‡0.363	‡0.827
ECRTM	**0.431**	**0.964**	**0.405**	**0.904**	**0.393**	**0.974**	**0.373**	**0.887**	**0.405**	**0.985**	**0.389**	**0.903**	**0.466**	**0.961**	0.416	**0.981**

Table 2: Topic quality of topic coherence (C_V) and topic diversity (TD) under 50 and 100 topics (K=50 and K=100). The best scores are in **bold**. ‡ means the gain of ECRTM is statistically significant at 0.05 level.

Model	20NG				IMDB				Yahoo Answer				AG News			
	K=50		K=100		K=50		K=100		K=50		K=100		K=50		K=100	
	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI
LDA	‡0.367	‡0.364	‡0.364	‡0.346	‡0.614	‡0.041	‡0.600	‡0.037	‡0.288	‡0.144	‡0.297	‡0.148	‡0.640	‡0.193	‡0.654	‡0.194
WLDA	‡0.233	‡0.157	‡0.292	‡0.207	‡0.589	‡0.011	‡0.602	‡0.013	‡0.255	‡0.084	‡0.303	‡0.127	‡0.580	‡0.151	‡0.653	‡0.188
DVAE	‡0.087	‡0.018	‡0.104	‡0.035	‡0.517	‡0.000	‡0.525	‡0.001	‡0.171	‡0.030	‡0.202	‡0.057	‡0.713	‡0.219	‡0.407	‡0.030
ETM	‡0.347	‡0.319	‡0.394	‡0.339	‡0.660	‡0.038	‡0.648	‡0.037	‡0.405	‡0.192	‡0.428	‡0.208	‡0.679	‡0.224	‡0.674	‡0.204
HyperMiner	‡0.433	‡0.405	‡0.454	‡0.386	‡0.655	‡0.046	‡0.641	‡0.032	‡0.456	‡0.237	‡0.448	‡0.222	‡0.730	‡0.276	‡0.679	‡0.200
NSTM	‡0.354	‡0.356	‡0.383	‡0.363	‡0.658	‡0.040	‡0.659	‡0.039	‡0.395	‡0.241	‡0.405	‡0.242	‡0.719	‡0.324	‡0.764	‡0.359
WeTe	‡0.268	‡0.304	‡0.338	‡0.348	‡0.587	‡0.031	‡0.589	‡0.025	‡0.389	‡0.252	‡0.444	‡0.269	‡0.641	‡0.268	‡0.699	‡0.271
ECRTM	**0.560**	**0.524**	**0.555**	**0.494**	**0.694**	**0.058**	**0.694**	**0.049**	**0.550**	**0.295**	**0.563**	**0.311**	**0.802**	**0.367**	**0.812**	**0.428**

Table 3: Document clustering of Purity and NMI under 50 and 100 topics (K=50 and K=100). The best scores are in **bold**. ‡ means the gain of ECRTM is statistically significant at 0.05 level.

topics for the above topic quality evaluation. We furthermore conduct **document clustering** experiments to evaluate doc-topic distribution quality with Purity and NMI following Zhao et al. (2021b); Wang et al. (2022).

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. Note that we cannot use it for document clustering since it cannot infer the doc-topic distributions. (iii) **DVAE** (Burkhardt & Kramer, 2019), Dirichlet VAE that approximates Dirichlet priors with rejection sampling; (iv) **WLDA** (Nan et al., 2019), a WAE-based topic model; (v) **ETM** (Dieng et al., 2020), a neural topic model which models the topic-word distribution matrix with word and topic embeddings; (vi) **HyperMiner** (Xu et al., 2022), using embeddings in the hyperbolic space to model topics. (vii) **NSTM** (Zhao et al., 2021b), using optimal transport distance between doc-topic distributions and documents to measure reconstruction error. (viii) **WeTe** (Wang et al., 2022), following NSTM and using conditional transport distance as reconstruction error.

4.2. Topic and Doc-Topic Distribution Quality

Table 2 reports the topic quality results concerning C_V and TD, and Table 3 summarizes doc-topic distribution quality results concerning Purity and NMI of document clustering. We have the following observations: **(i) ECRTM effectively addresses the topic collapsing issue and outperforms baselines in topic quality.** In Table 2, the much lower TD scores imply baselines generate repetitive topics and thus suffer from topic collapsing. As aforementioned, these repetitive topics are less useful for downstream tasks and damage the interpretability of topic models. In contrast, we see our ECRTM achieves significantly higher TD scores across all datasets and mostly the best C_V scores. We emphasize although the C_V of ECRTM is slightly higher than NSTM (0.389 v.s. 0.387) on Yahoo Answer, ECRTM completely outperforms on TD (0.903 v.s. 0.659). These results demonstrate that ECRTM produces more coherent and diverse topics than state-of-the-art baselines. These improvements are because our ECRTM makes topic embeddings away from each other and cover different semantics of word embeddings in the space instead of collapsing together as some baselines. **(ii) ECRTM surpasses baselines in inferring high-quality doc-topic dis-**

Model	20NG				Yahoo Answer			
	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD
DKM	‡0.510	‡0.471	0.448	‡0.577	‡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	‡0.504	‡0.446	0.461	‡0.548	‡0.498	‡0.262	0.435	‡0.608
ECRTM	0.560	0.524	0.431	0.964	0.550	0.295	0.405	0.985

Table 4: Ablation study. The terrible TD scores of DKM, DKM+Entropy and w/o ECR indicate topic collapsing still exists, making their high C_V less meaningful. In contrast, ECRTM achieves much higher TD with the best Purity and NMI. The C_V scores of ECRTM also outperforms state-of-the-art baselines (See Table 2). ‡ means the gain of ECRTM is statistically significant at 0.05 level.

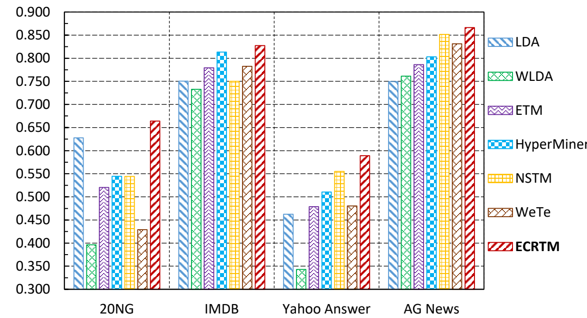


Figure 4: Text classification results of F1 scores. The improvements of ECRTM are all statistically significant at 0.05 level.

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
HyperMiner	one even end way little part character make never plot even end little way one plot character enough part make even seems fact enough plot end least character audience make
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	just like really bad good get one think see even man back gets goes two get takes house around away jokes jackson lisa predictable recycled wasted murphy writers williams rock
DKM	christmas disney musical songs bill timeless prince art rock holiday christmas santa childrens holiday betty age ann adult children toy fantasy christmas magic effects magical santa special holiday childrens child
DKM+Entropy	funny day physical semi ever way old due seen zone funny ever day old seen way physical due semi relationship funny ever day semi seen physical way old due psychological
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

Table 5: Case study: each row is the top 10 related words of a discovered topic. Repetitive words are underlined.

tributions. Table 3 shows 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 0.367 and 0.364. These manifest that ECRTM not only achieves higher-quality topics but also better doc-topic distributions as document representations. **See Appendices C to F for more experiments like robustness to the number of topics and visualization results.**

4.3. Ablation Study

We conduct ablation studies and show the necessity of our proposed Embedding Clustering Regularization (ECR). Specifically, we remove the ECR from 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 (DKM+Entropy, see Sec. 3.3). Note that we are the first to use DKM in topic modeling. Table 4 shows DKM, DKM+Entropy, and w/o ECR all suffer from topic collapsing as indicated by their much lower TD scores. Although they have high C_V, their terrible TD scores mean most topics are repetitive and less useful for downstream tasks, making their high C_V scores less meaningful (see examples in Sec. 4.5 for illustrations). Conversely, our ECRTM improves TD scores by a large margin and achieves the best document clustering performance with much higher Purity and NMI. This is because our ECR, as an effective regularization, can avoid the collapsing of topic embeddings while DKM, DKM+Entropy, and w/o ECR cannot. These results demonstrate our ECR is necessary to address the topic collapsing issue and achieve effective topic modeling performance.

Model	max-df=0.5				max-df=0.4				max-df=0.3				max-df=0.2				max-df=0.1			
	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI	C_V	TD
LDA	0.402	0.375	0.378	0.661	0.395	0.372	0.367	0.685	0.431	0.394	0.379	0.724	0.403	0.407	0.379	0.733	0.460	0.428	0.389	0.820
KM	/	/	0.264	0.193	/	/	0.277	0.260	/	/	0.270	0.197	/	/	0.267	0.240	/	/	0.260	0.203
WLDA	0.221	0.154	0.353	0.409	0.223	0.144	0.343	0.413	0.222	0.150	0.351	0.388	0.232	0.153	0.365	0.420	0.243	0.167	0.376	0.473
HyperMiner	0.465	0.429	0.360	0.600	0.430	0.398	0.356	0.583	0.445	0.425	0.363	0.599	0.457	0.434	0.369	0.577	0.495	0.444	0.375	0.708
ETM	0.365	0.330	0.384	0.736	0.367	0.301	0.367	0.735	0.354	0.307	0.372	0.735	0.367	0.307	0.379	0.760	0.370	0.319	0.376	0.844
NSTM	0.351	0.351	0.398	0.435	0.383	0.382	0.388	0.423	0.447	0.409	0.394	0.432	0.396	0.388	0.399	0.511	0.393	0.365	0.405	0.620
WeTe	0.381	0.422	0.380	0.944	0.299	0.349	0.385	0.961	0.311	0.387	0.371	0.923	0.326	0.395	0.384	0.957	0.376	0.450	0.394	0.957
ECRTM	**0.556**	**0.525**	**0.431**	**0.992**	**0.516**	**0.503**	**0.417**	**0.992**	**0.599**	**0.556**	**0.436**	**0.952**	**0.579**	**0.524**	**0.432**	**0.987**	**0.546**	**0.514**	**0.416**	**0.976**

Table 6: Influence of high-frequency words. Here max-df denotes the maximum document frequency of dataset pre-processing. A smaller max-df removes more high-frequency words. The best scores are in **bold**.

4.4. Text Classification

To evaluate extrinsically, we further conduct text classification experiments as downstream tasks. Specifically, we use the doc-topic distributions learned by topic models as document features and train SVMs to predict the class of each document. As reported in Figure 4, ECRTM significantly outperforms baseline models on all datasets. These results demonstrate that our ECRTM can be better utilized in the downstream classification tasks.

4.5. Case Study: Examples of Discovered Topics

For case study, Table 5 shows examples of discovered topics by different models from IMDB. We observe that ETM and NSTM both have highly uninformative and similar topics including common words like “just”, “like”, or “something”. HyperMiner generates repetitive topics with the words “one”, “even”, and “end” WeTe produces some less informative topics like “just like really bad good...”. DKM and DKM+Entropy also have repetitive topics with the words “christmas”, “holiday”, and “funny”. Accordingly, we observe that the topic collapsing issue commonly exists in these methods. These collapsed topics are uninformative and redundant, which are less useful for downstream applications 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 more coherent, such as the first topic with relevant words like “jackie”, “chan”, and “stunts”. Appendix G shows the full topics lists of models.

4.6. Influence of Dataset Pre-processing

As aforementioned in Sec. 3.2, we argue that topic collapsing results from the reconstruction error minimization on high-frequency words. This inspires us to ask: what if we carefully remove high-frequency words using reliable dataset pre-processing? Driven by this, we alter the maximum document frequency (max-df) to remove the high-frequency words when pre-processing the 20NG dataset. A smaller max-df removes more high-frequency words. From Table 6 we see that most baselines, such as LDA, ETM, and NSTM, reach higher TD scores under small max-df, indicating that topic collapsing is alleviated to some extent. But our model consistently outperforms all baselines on the topic quality and document clustering. These results empirically confirm our argument that the topic collapsing issue arises from reconstructing high-frequency words. Besides, these results verify one of our advantages: **our model requires no reliable pre-processing to achieve state-of-the-art performance**. This advantage is vital since the definition of reliable pre-processing is inconclusive: A large max-df may not remove any high-frequency words while a small one may remove most of the important words. More critically, brutally searching for reliable pre-processing is time-consuming and laborious. This advantage becomes more significant when meeting many large-scale datasets from various domains.

5. Conclusion

In this paper, we propose the novel Embedding Clustering Regularization Topic Model (ECRTM) to address the topic collapsing issue. ECRTM learns topics under the new Embedding Clustering Regularization that forces each topic embedding to be the center of a separately aggregated word embedding cluster. Extensive experiments demonstrate that ECRTM achieves effective neural topic modeling, successfully alleviates topic collapsing, and consistently achieves state-of-the-art performance in terms of producing high-quality topics and topic distributions of documents.

Acknowledgements

We thank all anonymous reviewers for their helpful comments. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme, AISG Award No: AISG2-TC-2022-005 and AISG Award No: AISG-100E-2019-046.

References

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

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

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

Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pp. 31–40, 2009.

Boyd-Graber, J. L., Hu, Y., Mimno, D., et al. *Applications of topic models*, volume 11. now Publishers Incorporated, 2017.

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

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

Card, D., Tan, C., and Smith, N. A. 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.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pp. 288–296, 2009.

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

Das, R., Zaheer, M., and Dyer, C. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 795–804, 2015.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Fard, M. M., Thonet, T., and Gaussier, E. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138:185–192, 2020.

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

Genevay, A., Dulac-Arnold, G., and Vert, J. Differentiable deep clustering with cluster size constraints. *CoRR*, abs/1910.09036, 2019. URL http://arxiv.org/abs/1910.09036.

Gupta, P., Chaudhary, Y., Buettner, F., and Schütze, H. Document informed neural autoregressive topic models with distributional prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6505–6512, 2019.

Hennig, P., Stern, D., Herbrich, R., and Graepel, T. Kernel topic models. In *Artificial Intelligence and Statistics*, pp. 511–519, 2012.

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

Hsu, C.-C. and Lin, C.-W. 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.

Huang, P., Huang, Y., Wang, W., and Wang, L. Deep embedding network for clustering. In *2014 22nd International conference on pattern recognition*, pp. 1532–1537. IEEE, 2014.

Huynh, V., Zhao, H., and Phung, D. Otlda: A geometry-aware optimal transport approach for topic modeling. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (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.

Kim, H., Choo, J., Kim, J., Reddy, C. K., and Park, H. 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.

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

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

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

Ma, Z., Sun, A., Yuan, Q., and Cong, G. 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.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. 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.

McAuley, J. and Leskovec, J. 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.

Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., and Han, J. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pp. 2121–2132, 2020.

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

Miao, Y., Grefenstette, E., and Blunsom, P. 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.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. 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.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 262–272. Association for Computational Linguistics, 2011.

Nan, F., Ding, R., Nallapati, R., and Xiang, B. 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.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. 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.

Nguyen, T. and Luu, A. T. Contrastive learning for neural topic model. *Advances in Neural Information Processing Systems*, 34, 2021.

Pennington, J., Socher, R., and 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.

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

Piantadosi, S. T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

Reed, W. J. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *In Proceedings ofthe 31th International Conference on Machine Learning*, 2014.

Röder, M., Both, A., and Hinneburg, A. 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.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. 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.

Sia, S., Dalmia, A., and Mielke, S. J. 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.

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

Song, C., Liu, F., Huang, Y., Wang, L., and Tan, T. Autoencoder 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.

Srivastava, A. and Sutton, C. 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.

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

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

Wallach, H., Mimno, D., and McCallum, A. Rethinking lda: Why priors matter. *Advances in neural information processing systems*, 22, 2009.

Wang, C. and Blei, D. M. 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.

Wang, D., Guo, D., Zhao, H., Zhang, H., Tanwisuth, K., Chen, B., and Zhou, M. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=IYMuTbGzjFU.

Wang, X., McCallum, A., and Wei, X. 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.

Wu, X. and Li, C. Short Text Topic Modeling with Flexible Word Patterns. In *International Joint Conference on Neural Networks*, 2019.

Wu, X., Li, C., Zhu, Y., and Miao, Y. Learning Multilingual Topics with Neural Variational Inference. In *International Conference on Natural Language Processing and Chinese Computing*, 2020a.

Wu, X., Li, C., Zhu, Y., and Miao, Y. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1772–1782, Online, November 2020b.

Wu, X., Li, C., and Miao, Y. Discovering topics in long-tailed corpora with causal intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 175–185, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.15. URL https://aclanthology.org/2021.findings-acl.15.

Wu, X., Luu, A. T., and Dong, X. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2748–2760, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.176.

Wu, X., Dong, X., Nguyen, T., Liu, C., Pan, L., and Luu, A. T. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. *arXiv preprint arXiv:2304.03544*, 2023.

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

Xu, Y., Wang, D., Chen, B., Lu, R., Duan, Z., and Zhou, M. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31557–31570. Curran Associates, Inc., 2022.

Yan, X., Guo, J., Lan, Y., and Cheng, X. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456. ACM, 2013a.

Yan, X., Guo, J., Liu, S., Cheng, X., and Wang, Y. 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, 2013b.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pp. 3861–3870. PMLR, 2017.

Yao, E., Zheng, G., Jin, O., Bao, S., Chen, K., Su, Z., and Yu, Y. Probabilistic text modeling with orthogonalized topics. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, pp. 907–910, 2014.

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

Zhang, Z., Fang, M., Chen, L., and Namazi-Rad, M.-R. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. *arXiv preprint arXiv:2204.09874*, 2022.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*, 2021a.

Zhao, H., Phung, D., Huynh, V., Le, T., and Buntine, W. L. 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.

Model	K=10		K=20		K=30		K=40		K=60		K=70		K=80		K=90	
	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD	C_V	TD
LDA	0.376	0.627	0.384	0.620	0.377	0.678	0.390	0.697	0.373	0.674	0.385	0.590	0.381	0.645	0.380	0.656
KM	0.208	0.207	0.230	0.180	0.247	0.151	0.255	0.217	0.269	0.219	0.282	0.256	0.300	0.321	0.289	0.300
WLDA	0.354	0.533	0.354	0.443	0.360	0.389	0.357	0.430	0.361	0.343	0.373	0.335	0.371	0.345	0.375	0.287
ETM	0.380	0.820	0.372	0.763	0.373	0.744	0.383	0.698	0.367	0.672	0.377	0.667	0.377	0.663	0.372	0.629
HyperMiner	0.361	0.800	0.377	0.750	0.373	0.671	0.378	0.665	0.379	0.590	0.380	0.533	0.373	0.516	0.369	0.454
NSTM	0.397	0.600	0.381	0.487	0.385	0.391	0.389	0.418	0.392	0.444	0.396	0.520	0.400	0.468	0.386	0.450
WeTe	0.422	**1.000**	0.380	0.980	0.387	0.980	0.388	0.978	0.378	0.948	0.368	0.879	0.355	0.800	0.349	0.754
ECRTM	**0.487**	**1.000**	**0.454**	**1.000**	**0.437**	**1.000**	**0.435**	**0.993**	**0.413**	**0.993**	**0.405**	**0.910**	**0.410**	**0.957**	**0.402**	**0.906**

Table 7: Topic quality of coherence (C_V) and diversity (TD) under topic number $K = 10, 20, 30, 40, 60, 70, 80, 90$. The best scores are in **bold**.

Model	K=10		K=20		K=30		K=40		K=60		K=70		K=80		K=90	
	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI	Purity	NMI
LDA	0.295	0.408	0.340	0.396	0.347	0.375	0.368	0.356	0.354	0.352	0.399	0.368	0.378	0.352	0.389	0.359
WLDA	0.174	0.119	0.194	0.124	0.223	0.152	0.238	0.161	0.260	0.176	0.272	0.186	0.252	0.180	0.281	0.199
ETM	0.183	0.274	0.275	0.307	0.307	0.288	0.331	0.281	0.351	0.291	0.340	0.302	0.379	0.330	0.407	0.349
HyperMiner	0.240	0.299	0.338	0.390	0.416	0.421	0.407	0.389	0.478	0.422	0.461	0.408	0.468	0.396	0.446	0.384
NSTM	0.228	0.284	0.295	0.327	0.355	0.373	0.349	0.349	0.362	0.353	0.357	0.344	0.351	0.365	0.376	0.354
WeTe	0.055	0.004	0.119	0.150	0.197	0.244	0.252	0.317	0.281	0.332	0.384	0.421	0.313	0.331	0.302	0.311
ECRTM	**0.390**	**0.485**	**0.373**	**0.420**	**0.463**	**0.435**	**0.462**	**0.426**	**0.554**	**0.522**	**0.559**	**0.498**	**0.581**	**0.506**	**0.564**	**0.497**

Table 8: Document clustering of Purity and NMI under topic number K=10, 20, 30, 40, 60, 70, 80, 90. The best scores are in **bold**.

Dataset	#docs	Vocabulary Size	Average Length	#labels
20NG	18,846	5,000	110.5	20
IMDB	50,000	5,000	95.0	2
Yahoo Answer	12,500	5,000	35.4	10
AG News	12,500	5,000	20.1	4

Table 9: Statistics of datasets after pre-processing.

Model	20NG	IMDB	Yahoo Answer	AG News
LDA	2044.6	2482.8	4637.1	9951.1
DVAE	2045.0	1996.3	3258.3	1851.3
ETM	2113.4	1911.8	4653.4	1518.7
ECRTM	**1896.9**	**1830.8**	**3244.0**	**1436.1**

Table 10: Perplexity results. The best is in **bold** (lower is better).

A. Dataset

We follow the dataset pre-processing steps of (Card et al., 2018): (1) tokenize documents and convert to lowercase; (2) remove punctuation; (3) remove tokens that include numbers; (4) remove tokens less than 3 characters; (5) remove stop words. The statistics of pre-processed datasets are reported in Table 9.

B. Model Implementation

For pre-trained word embeddings, we employ 200-dimensional GloVe (Pennington et al., 2014) [3]. 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 our 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. We set α as 1.0 following Card et al. (2018). Our encoder network is the same as Srivastava & Sutton (2017); Wu et al. (2020a;b): 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 (Kingma & Ba, 2014) to optimize model parameters. See other implementation details in our code.

[3] https://nlp.stanford.edu/projects/glove/

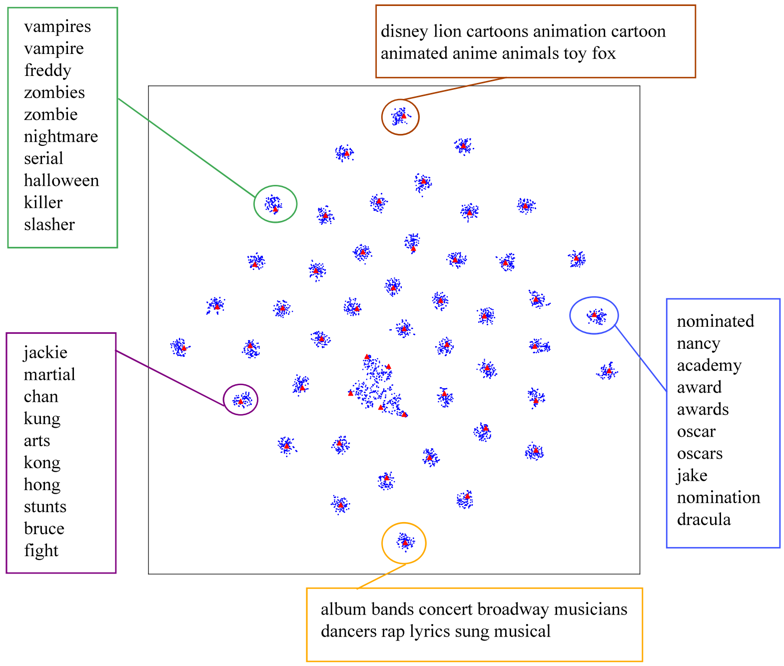
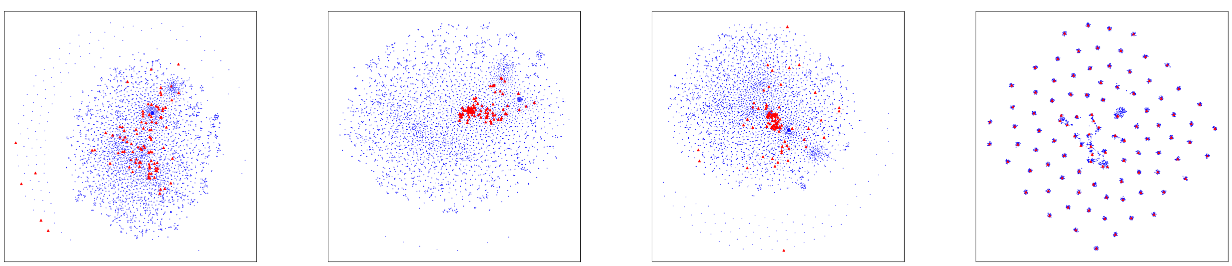


Figure 5: Annotations of top words of discovered topics in the semantic space.



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

Figure 6: 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 by forcing each topic embedding to be the center of a separately aggregated word embedding cluster.

C. Perplexity Results

We report the perplexity results in Table 10. Here we do not include some neural topic models (WLDA, NSTM, WeTe) as they are inapplicable to the perplexity approximation with ELBO (See explanations in Miao et al. (2016); Srivastava & Sutton (2017); Nan et al. (2019); Zhao et al. (2021a)). Table 10 shows our ECRTM also achieves the best perplexity results (lower is better).

D. Robustness to the Number of Topics

Besides the aforementioned results under $K = 50, 100$ (Tables 2 and 3), we also experiment under $K = 10, 20, 30, 40, 60, 70, 80, 90$ on 20NG to verify the robustness of our method. As shown in Tables 7 and 8, we see that ECRTM consistently outperforms baseline models in terms of both topic quality and clustering. These show that the performance improvements of our ECRTM are robust to the number of topics.

E. Comparison of Coherence Metrics

Röder et al. (2015) have empirically shown that C_V is a better coherence metric which has better consistency with human judgment than traditional metrics like NPMI, UCI, and UMass (Bouma, 2009; Chang et al., 2009; Newman et al., 2010; Mimno et al., 2011). We also confirm this argument in our experiments: we find NPMI, UCI, and UMass tend to

NPMI	UCI	UMass	C_V	Top words of topics
0.069	0.746	-0.900	0.272	**Topic#1:** way come actually make yet example even fact though rather
0.065	0.615	-0.967	0.272	**Topic#2:** fact even indeed though kind something way always actually things
0.065	0.628	-3.367	0.255	**Topic#3:** really pretty something seem seems quite things nothing thing ridiculous
-0.076	-2.660	-8.196	0.497	**Topic#4:** vampires vampire freddy zombies zombie nightmare serial halloween killer slasher
-0.075	-2.678	-9.032	0.476	**Topic#5:** disney lion cartoons animation cartoon animated anime animals toy fox
-0.114	-3.367	-6.120	0.452	**Topic#6:** jackie martial chan kung arts kong hong stunts bruce fight

Table 11: Comparison of coherence metrics. C_V gives high scores to coherent topics (Topic#4-6) while traditional metrics (NPMI, UCI, UMass) assigns high scores to less informative topics (Topic#1-3).

give higher scores to trivial and less informative topics. For example, we see in Table 11, Topic#4-6 are more coherent and informative than Topic#1-3. But unfortunately, NPMI, UCI, and UMass give much higher scores to Topic#1-3 instead of Topic#4-6. In contrast, C_V is more reasonable as it gives higher scores to Topic#4-6. Note that it is incorrect to directly compare a NPMI/UCI/UMass score with a C_V score because they are in different scales.

F. Visualization of Embedding Space

We visualize the learned topic and word embeddings with t-SNE (van der Maaten & Hinton, 2008) **under 100 topics** (Figure 1 is under 50 topics). Figure 6 shows while the topic embeddings mostly collapse together in the state-of-the-art baselines, our ECRTM avoids the collapsing of topic embeddings by forcing each topic embedding to be the center of a separately aggregated word embedding cluster. This illustrates that our ECR also works effectively under a larger number of topics.

We furthermore annotate the semantic embedding space with the top related words of discovered topics by ECRTM under K=50 as shown in Figure 5. We see that each word embedding cluster represents a diverse and coherent topic. This shows that our ECRTM effectively clusters the embeddings of coherent words.

G. Full Lists of Discovered Topics

Below are the discovered topics of different models from IMDB under 50 topics (K=50).

ETM

```
Topic#1:  director etc dialog cult ben joe steve mans todays bruce
Topic#2:  killer police kill killed murder car head shot killing body
Topic#3:  instead rather put mind together despite come beyond without alone
Topic#4:  series first later one time also since made set years
Topic#5:  art rather use special little like sci king bit style
Topic#6:  dark fantasy earth full begin world open end setting set
Topic#7:  family father young wife son mother child life children daughter
Topic#8:  show shows episode watching see watch episodes television star dvd
Topic#9:  boring entertaining pointless silly dull humour pretentious sub unfunny entertained
Topic#10:  like better good especially end look much done way just
Topic#11:  war world history people documentary american life society german soldiers
Topic#12:  film films cinema directors filmed hollywood filmmakers cinematography welles noir
Topic#13:  new one way time world different place two live day
Topic#14:  like just one way made much times really even feel
Topic#15:  music musical song songs dance stage oscar singing voice career
Topic#16:  even one though also far way still made yet better
Topic#17:  part like end also along two short lines now line
Topic#18:  self somewhat sense seems narrative character sexual becomes often person
Topic#19:  story book great perfect gives comes work novel excellent brilliant
Topic#20:  supposed keep get wanted got trying never kept going turn
Topic#21:  people just really know think say maybe something things understand
Topic#22:  john michael james robert richard paul george jack cast played
Topic#23:  scene scenes sex plot slow violence blood boring twist nudity
Topic#24:  action man death gets fight hero back another comes face
Topic#25:  plot original poor pretty unfortunately decent terrible nothing budget completely
Topic#26:  story also makes interesting many stories however quite make style
Topic#27:  great really good see recommend fan liked definitely thought watch
Topic#28:  seen ever one worst years saw see beginning time remember
Topic#29:  one time make give anyone made take get making another
Topic#30:  time minutes just long hour half back night couple looking
Topic#31:  house goes starts gets looks red night tries takes strange
Topic#32:  love life heart beautiful human loved lives wonderful amazing world
Topic#33:  first one ending time two end second just made last
Topic#34:  best role actor performance played also good performances actors play
Topic#35:  characters character plot director dialogue film real drama development acted
Topic#36:  actors cast good acting throughout script moments story film written
Topic#37:  really thing good guy watch movie whole yes want waste
Topic#38:  course much seems audience quite less picture overall bit opinion
Topic#39:  man young men women two woman town small female local
Topic#40:  like high school game video one quality camera sound shot
Topic#41:  bad even acting like just awful made stupid worse crap
Topic#42:  old little girl get kids girls boy kid cool year
Topic#43:  movie movies watch watching actors people theater theaters disaster entertainment
Topic#44:  make see sure say think need really looks know good
Topic#45:  version american truly classic japanese british era french cartoon tale
Topic#46:  funny comedy fun laugh humor jokes watch hilarious comic imagine
Topic#47:  one like around sort looking kind good main look just
Topic#48:  actually real black just white believe mean see either know
Topic#49:  horror effects low dvd english movies films zombie dead genre
Topic#50:  fact reason actually even never simply none however obvious nothing
```

HyperMiner

```
Topic#1:  jack musical oscar song songs george dance jane mary scott
Topic#2:  just people like think know something see get say want
Topic#3:  effort attempts lacks poor merely intended lacking whilst grace appeal
Topic#4:  one even end way little part character make never plot
Topic#5:  also quite interesting however rather many much although bit though
Topic#6:  one great seen ever best see watch time every enjoy
Topic#7:  even one end way little part character make plot never
Topic#8:  just like people know something think say see want get
Topic#9:  life love beautiful real heart TRUE romantic lives romance dream
Topic#10:  even one end part little character way make never point
Topic#11:  film films made cinema making silent festival director makers filmmakers
Topic#12:  scene get gets guy girl around back away head getting
Topic#13:  movie movies bad made watch make acting watching even plot
Topic#14:  see saw thought watched dvd watching got video felt went
Topic#15:  good like really just better lot much look pretty nice
Topic#16:  bad horror worst awful terrible acting waste budget worse low
Topic#17:  police crime cop gun soldiers gang prison church agent charlie
Topic#18:  even end way character seems make never little part fact
Topic#19:  funny comedy fun kids school laugh humor jokes hilarious christmas
Topic#20:  just people know like think something see say get want
Topic#21:  even seems fact enough plot end least character audience make
Topic#22:  show series shows episode episodes television season pilot trek writers
Topic#23:  car game red dog cat camp van steve chase eye
Topic#24:  one first two new time another second star also world
Topic#25:  effects earth monster sci space battle island match computer adventure
Topic#26:  one great seen ever best see watch time every fan
Topic#27:  man takes find woman help place small comes becomes along
Topic#28:  also many however quite much interesting rather although though bit
Topic#29:  scene get gets around guy girl sex getting back away
Topic#30:  cast role performance actor john excellent play played performances actors
Topic#31:  just like people know think see get something say going
Topic#32:  film films made cinema making silent director independent makers festival
Topic#33:  scenes action dark slow genre fight opening scene sequence violence
Topic#34:  scene get gets guy girl around back sex guys getting
Topic#35:  one first two time new another second also star world
Topic#36:  even end way one little character make plot enough part
Topic#37:  director work camera production music script sound writer shot direction
Topic#38:  human art french world nature reality powerful images experience deep
Topic#39:  war american black white history world country english british documentary
Topic#40:  young family wife father boy son mother children child daughter
Topic#41:  one great seen ever see best watch time every fan
Topic#42:  one great seen ever best see watch time fan every
Topic#43:  old years still now last time year three long back
Topic#44:  death dead evil blood killer house kill night murder killed
Topic#45:  even end little way one plot character enough part make
Topic#46:  lack attempt self premise dull flat fails failed unfortunately pretentious
Topic#47:  story characters original book version read based stories character king
Topic#48:  good like really just better lot much look pretty acting
Topic#49:  cast role performance actor john excellent played play actors performances
Topic#50:  saw see thought dvd watched watching felt got went video
```

NSTM

```
Topic#1:  miller got smith moore just johnson really know davis think
Topic#2:  even one though time way just come fact much make
Topic#3:  know just really going think something come maybe get even
Topic#4:  one another man murder even wanted others death just taken
Topic#5:  movie like monster something actually just movies come thing kind
Topic#6:  one just along now part way another come time around
Topic#7:  just get going come one know even really something way
Topic#8:  one first time also though best another even although came
Topic#9:  sense kind something love really feel thing feeling nothing sort
Topic#10:  movie movies film films just really best like something thing
Topic#11:  just like really come going get even maybe something good
Topic#12:  really movie thing something just things maybe good stuff kind
Topic#13:  interesting movie something funny kind really quite wonderful things fun
Topic#14:  just going really get something maybe come know even thing
Topic#15:  one time first movie best though even just like also
Topic#16:  best one smith moore miller time first just davis james
Topic#17:  movie film just even best though films actually fact really
Topic#18:  just show even come time one good really going know
Topic#19:  wonderful amazing terrific good really fantastic best something thing pretty
Topic#20:  wearing wore wear dress look dressed wears clothes worn shirt
Topic#21:  even just come way going get make time though one
Topic#22:  really something think know thing maybe just things going good
Topic#23:  movie film best fact one though example even story life
Topic#24:  music best musical songs movie like featured one song playing
Topic#25:  really something seems pretty quite things seem thing nothing think
Topic#26:  goes takes tells gets comes finds makes knows everyone happens
Topic#27:  just even really something come going like actually things get
Topic#28:  just one even something come way really like always good
Topic#29:  funny silly movie stuff amusing scary fun hilarious cheesy boring
Topic#30:  even though much make way just come actually fact one
Topic#31:  just even come one way good really though going something
Topic#32:  even fact though way come actually make yet indeed something
Topic#33:  just one even come time way going coming another get
Topic#34:  something really even just things good actually always way kind
Topic#35:  one even time now though come came last another also
Topic#36:  best movie film one good films actor like movies time
Topic#37:  movie film movies films comedy drama best hollywood starring story
Topic#38:  mother daughter wife sister friend husband married couple actress love
Topic#39:  fact even example though way one rather much indeed life
Topic#40:  nose eyes hand just mouth legs fingers neck teeth like
Topic#41:  really just something think going know things maybe even thing
Topic#42:  even come know just fact think something way really actually
Topic#43:  one even just come now time life mother know though
Topic#44:  just like inside get even look come one everything away
Topic#45:  really maybe thing know something think things just going everybody
Topic#46:  one just another came time back now come went got
Topic#47:  fact even something really though always way actually things kind
Topic#48:  time first one just last second next came coming play
Topic#49:  horrible awful terrible horrific thing horrifying frightening kind shocking really
Topic#50:  film movie films movies best directed drama feature picture though
```

WeTe

Topic#1: tonight fifth terrific kings loser wow fourth lucky bang grabs
Topic#2: australia progress reached secondly environment suspects interests scores continued press
Topic#3: profanity striking refuse stress sue complain survived fatal contact cracking
Topic#4: located population neighborhood nation distance owns cox traffic centered bell
Topic#5: graphics wholly puzzle map chapter attraction medium reader composed edition
Topic#6: acclaimed exquisite gothic vivid splendid stark sublime lively photographer literary
Topic#7: theaters cinemas preview trailers ratings extras studios rental par nyc
Topic#8: journalist calls debate complaint behaviour flawed defense civil charge innocence
Topic#9: discovering pursuit rid junk disappears cheating pretending petty discovers saving
Topic#10: john big new star match city james game george stars
Topic#11: predator heartbreaking terrifying bravo suspenseful intrigue paranoia menace unsettling mafia
Topic#12: story characters life character people way real sense love much
Topic#13: pal moody greek trend global corporate depression combat transformed sin
Topic#14: shaky hung gray shine rough heights staring screens clad casts
Topic#15: blacks teams merit races thru selection junior earned ranks groups
Topic#16: interestingly damned awfully astounding unreal incidentally screwed rendered alas instantly
Topic#17: original version book classic read novel adaptation written king sequel
Topic#18: adding sticking thread added process easier hang repeating suspend pieces
Topic#19: recommendation months raped policeman restored retired morning month weeks six
Topic#20: sorry understands regret lately teenager fond disliked worry unhappy troubles
Topic#21: show first years series time see saw since still now
Topic#22: records bollywood pulp futuristic romp bars circus punk hardcore gems
Topic#23: huh gotta cried swear guessed shouting shake fooled dude yelling
Topic#24: builds adds reaches agrees teaches introduces marries explains resembles threatens
Topic#25: rat tank burn duck trees dirt rabbit tree snake burning
Topic#26: film one films scenes director also time story plot even
Topic#27: brad ron betty matt dan ryan flynn glover anderson ann
Topic#28: war american world black white documentary history america people political
Topic#29: family young kids school father old girl children child mother
Topic#30: music musical songs song dance voice dancing singing rock stage
Topic#31: familys anyones hitchcocks wouldve shouldve everyones couldve wifes itll expect
Topic#32: great best good role cast comedy actor love character actors
Topic#33: active musician venture learning overcome teaching taught remained accomplished concepts
Topic#34: sport gross caliber definition mere waves design games rip aforementioned
Topic#35: women sex girls woman female men violence sexual gay scenes
Topic#36: killer death murder police thriller cop crime kill mystery michael
Topic#37: abc holiday aired eve midnight remake introduce mtv broadcast began
Topic#38: movie movies watch watching acting plot story scenes wifes familys
Topic#39: just like really bad good get one think see even
Topic#40: man back gets goes two get takes house around away
Topic#41: alcoholic wine cruise vegas serving con businessman beverly bent california
Topic#42: tech korean victory capital temple wwii riot seconds empire dragon
Topic#43: recorded guitar album tracks reviewer recording blues noted singers sung
Topic#44: horror action effects special budget evil gore blood low fight
Topic#45: hoot tarantino bergman dracula buster creator mario olds hack buff
Topic#46: origin cultural versus conscience roots reunion politics christ divorce dignity
Topic#47: confident keen tad flair shy lyrics smart impressed nod calm
Topic#48: shoddy tasteless formulaic imaginable inane unfunny tripe overacting yawn incompetent
Topic#49: doom beware compassion foul mayhem rides subtlety cue awe wicked
Topic#50: simultaneously readers mute dimension complexity critical derivative essence perspective account

DKM

```
Topic#1:  cooper india family victor town indian dorothy uncle alice jake
Topic#2:  animated animation disney voiced lion bugs batman cartoon cartoons anime
Topic#3:  christmas disney musical songs bill timeless prince art rock loved
Topic#4:  gags gag footage channel jokes television pilot nostalgic smoking comedy
Topic#5:  rock magazine christian chris daddy roger access page jesus school
Topic#6:  jennifer eva comedy jokes genre plot humor sex nicole pie
Topic#7:  christ games jesus religion bible game freeman christian theory vampires
Topic#8:  eddie woody allen murphy comedians comedian keaton funnier sandra williams
Topic#9:  noises father fear heroine summer toys humor adolescent premise trick
Topic#10:  wedding prince grace betty affair sally marry opera secretary marriage
Topic#11:  winner neil star oscar debut hit screenplay jeff ben stars
Topic#12:  politically political racist garbage black wing jokes free foul stereotype
Topic#13:  children gay baby garbage andy virgin parents fox offensive abuse
Topic#14:  jokes jackson lisa predictable recycled wasted murphy writers williams rock
Topic#15:  waste crap worst costs garbage horrible wasting ashamed sucks pile
Topic#16:  erotic sexuality nudity explicit nude porn lesbian sexual sex photos
Topic#17:  seasons season episodes episode abc trek show aired sitcom series
Topic#18:  delightful parker grant gentle comedies henry witty grim delight arthur
Topic#19:  white black dated english clothing costumes costume queen period heroine
Topic#20:  halloween horror freddy scares slasher carpenter eerie haunted gory scary
Topic#21:  cried ned meryl streep charlie cry touched heartbreaking dan warming
Topic#22:  seagal abc jet tripe hardy stan martial heist arts drivel
Topic#23:  andrews noir detective eastwood harry hopper investigation fbi cop criminals
Topic#24:  lynch art imagery noir poetry images waters landscapes poetic symbolic
Topic#25:  moore female soap arthur daughter island butler copy nurse pregnant
Topic#26:  laugh sam winner comedy loser cop funniest golden hilarious simon
Topic#27:  bollywood indian india hudson khan soap dialogs indians dialogues opera
Topic#28:  santa homeless andrews truck airport car husband security mom wakes
Topic#29:  spike match angle baseball tag indians hudson stewart ring flynn
Topic#30:  sinatra musicals kelly powell mgm broadway musical dance numbers gene
Topic#31:  christmas santa childrens holiday betty age ann adult children toy
Topic#32:  robin don batman hood villains burt pacino dick delivery adam
Topic#33:  adaptation emma jane adaptations timothy bbc novel kenneth book bates
Topic#34:  sons philip son daniel father mother jewish hoffman norman davis
Topic#35:  cars rent scared channels vhs lol expecting cage renting cried
Topic#36:  woody disjointed pacing distracting uninteresting development leonard tension subplots hour
Topic#37:  trek alien planet superman science scientists invisible aliens outer scientist
Topic#38:  fantasy christmas magic effects magical santa special holiday childrens child
Topic#39:  gay blah gray spike indulgent lesbian jerk insulting rape racist
Topic#40:  martial arts kung jet hong chan kong ninja jackie sword
Topic#41:  zombies zombie werewolf vampires monster creature dinosaur rubber shark topless
Topic#42:  indie chaplin festival budget pleasantly independent jake distribution spike victor
Topic#43:  computer plot imaginative tech sub slasher budget holes video jennifer
Topic#44:  low code monkey rental store computer scientist laughs budget executive
Topic#45:  drew daughter adult father adults sports parents younger teen kids
Topic#46:  propaganda jews hitler germans nazis soviet nazi civil americans jewish
Topic#47:  baby jokes daughter humor amusing unfunny silly cute comedy funny
Topic#48:  bands metal album punk kids tap band santa nerd school
Topic#49:  bollywood manager eddie rap van store dance tap singer choreography
Topic#50:  eddie album nominated bands concert vhs murphy jackie awards oscars
```

DKM+Entropy

```
Topic#1:  zone european even choppy showing shows evening shy china side
Topic#2:  sincerely workers warn thank yes ever seen old way semi
Topic#3:  finally morgan drug tape tap tank drugs remember remind reminded
Topic#4:  laughter distracting discussion social disbelief disappoint moronic motivation dinosaurs dimension
Topic#5:  bin camera hollywoods couldve amazingly ambition pacing paced titled awe
Topic#6:  reviewers spend spoke got split gore gordon motion gone spike
Topic#7:  expected opened entry sea canada june north installment ticket april
Topic#8:  zone period overcome overlooked pace pacino discovering painful painfully paper
Topic#9:  naked cousin raped fat neo crazed fans twins indian indians
Topic#10:  abandoned tons discovers spoiled died die diane rat near nearby
Topic#11:  clothing fitting trained amazingly robot borrowed masterpiece primarily heavily comic
Topic#12:  overwhelming disjointed heartfelt flair understandable captures sincere angst spirited dreck
Topic#13:  buff dude like saw survivor genuine start estate saving ever
Topic#14:  physical day semi ever way old due funny seen harm
Topic#15:  zone insipid interestingly interesting interest single intentions intensity intelligence insulting
Topic#16:  roof practically underneath tame colour albeit colors shorter color badly
Topic#17:  episode identify corruption costs project ignored promise ignore couple couples
Topic#18:  getting spring dollars stevens steven message stephen step doctors mickey
Topic#19:  rip dogs images discovered chinese illegal objects baseball possibly display
Topic#20:  picked patients sandra hour hours howard boat human continually sarah
Topic#21:  zone brilliance principal prior brief describing prize description process produced
Topic#22:  urge decision rushed agree navy agreed agrees ahead aid defense
Topic#23:  won japan spain union april italy south bank north entry
Topic#24:  homeless womens disturbed helps cares compassion souls incompetent caring ashamed
Topic#25:  episode identify corruption costs project ignored promise ignore couple couples
Topic#26:  improved lacked relative reflection reflect stress corporate respect lighting supporting
Topic#27:  zone insipid interestingly interesting interest single intentions intensity intelligence insulting
Topic#28:  checked growing cole come comes hal gun guessed study comments
Topic#29:  rocket rescue slice mates burns strong stronger strongest bus respective
Topic#30:  semi seen due old way physical day ever funny sentimental
Topic#31:  benefit buying paid guarantee suits patricia wealthy pay paying pays
Topic#32:  cult japanese killers sympathy kill kevin justin josh joins jimmy
Topic#33:  zone playing physical picking piece battles credited place places plane
Topic#34:  canada lower april june entry north installment sea ticket opened
Topic#35:  struggle americans note notch barry barrel nostalgic task screen tape
Topic#36:  funny day physical semi ever way old due seen zone
Topic#37:  zone place phil philosophical philosophy class claire phrase physical claims
Topic#38:  invasion army wwii opened june bank join third next fourth
Topic#39:  way day ever old due seen funny physical semi person
Topic#40:  positive crisis elements painful serial television effects seen two causes
Topic#41:  gas plenty receives dislike average regarding treatment household safety concern
Topic#42:  zone honesty holy hollywood holiday holds shared hoffman hitting hits
Topic#43:  system college supported identical skip trial mentioned furthermore actual half
Topic#44:  funny ever day old seen way physical due semi relationship
Topic#45:  funny ever day semi seen physical way old due psychological
Topic#46:  stuff cried intentionally technical innocence anyway tender anything anyone incident
Topic#47:  compared conviction howard ted technology sci human officers office hundred
Topic#48:  tight understand positive came television seen standard win like semi
Topic#49:  ever semi due old funny physical day way seen solution
Topic#50:  nominated pass nod nomination suspect willing joined warner jean stanley
```

ECRTM

Topic#1: students school teacher student sean high specially class texas shelf
Topic#2: disney lion cartoons animation cartoon animated anime animals toy fox
Topic#3: dimensional merely clumsy unpleasant draw potentially consistent handed develop wholly
Topic#4: santa christmas children kids adults adult child parents relax age
Topic#5: stories york season match lifetime davis tony currently respected episodes
Topic#6: wars alien burton dinosaurs outer graphics trek futuristic aliens sci
Topic#7: terrible costs horrible renting sucks awful avoid rented sounded rent
Topic#8: budget low values stinks producing violence gratuitous blast spare lighting
Topic#9: seasons abc episode aired season show episodes network program television
Topic#10: eastwood andrews hopper fbi westerns investigation policeman crime investigating showdown
Topic#11: worst superman hats ever les shorter guitar banned maniac mates
Topic#12: funniest comedies comedy laugh laughing dan black mario white jokes
Topic#13: japanese japan russian wrestling reynolds kim biased marketing industry vincent
Topic#14: album bands concert broadway musicians dancers rap lyrics sung musical
Topic#15: bates adams purchased ann library lasted listed quinn map native
Topic#16: jackie martial chan kung arts kong hong stunts bruce fight
Topic#17: cars swedish car thumbs files cop airport sappy jet tomorrow
Topic#18: bat stuart angela shouldve flynn hoot plague trailer abysmal bath
Topic#19: sucked crappy stupid twins monkey stupidity darn horrid idiotic puppet
Topic#20: vhs bought copy tape rental dvd store dvds video cable
Topic#21: dad son jackson jack father hotel god saved king thinks
Topic#22: games game bond victor chris germany reunion mexico french marie
Topic#23: martha familys dysfunctional cowboy cope illness daniel bergman financial property
Topic#24: titanic waste spike flop dreck advise someday blockbuster junk taylor
Topic#25: grim page gentle timothy understated magnificent captures poignant passionate debut
Topic#26: development unlikeable drew roy main pitt character implausible descent believable
Topic#27: sequel remake original beginning credits improved van scene missed spoilers
Topic#28: ted lou chaplin bridges mel warren russell sandra butler elizabeth
Topic#29: football teams randy kicked jeff ensues airplane thugs roof bus
Topic#30: woody walken jerry kelly hanks allen sinatra tom dances musicals
Topic#31: lynch ireland bang unusual worthwhile rabbit tops twist quotes temple
Topic#32: excited waters comments reading expectations yelling book urge disturbed sticking
Topic#33: excellent fantastic recommend performance highly job brilliant amazing enjoyed definitely
Topic#34: vampires vampire freddy zombies zombie nightmare serial halloween killer slasher
Topic#35: christ christian religious faith theory interviews media holy intellectual studying
Topic#36: jane emma novel adaptation novels book version versions faithful books
Topic#37: freeman justin nicole holly carrie annie austin glover btw matthew
Topic#38: eddie murphy unfunny funnier tacky clown humour yawn gags tripe
Topic#39: artistic medium landscape breathtaking painting imaginative poetry technique contemporary movements
Topic#40: demons carpenter spooky eerie sleaze myers gothic karen hammer patients
Topic#41: rubber cabin ninja rat barrel cave corpses splatter lake predator
Topic#42: festival indie harry welles films buffs film stan lucy gay
Topic#43: mean maybe sorry hate honestly understand saying else noises someone
Topic#44: hitler germans soviet civil war wwii russia fought union holocaust
Topic#45: half improvement stretch ford hour limit respectable swimming covers portion
Topic#46: heaven cage bravo segment earth robot vegas science planet drove
Topic#47: wasting existent amateur mtv choppy lowest asleep shaky lifeless stores
Topic#48: nominated nancy academy award awards oscar oscars jake nomination dracula
Topic#49: virgin women andy sex sarah male diane kate woman boyfriend
Topic#50: touched bollywood sadness cried joy feelings emotion inspiring relationships warm