Neural Topic Model Aligning with Pre-trained Clustering and Optimal Transport

Anonymous submission

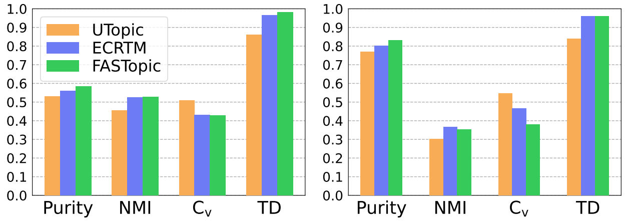
Abstract

Recent research on VAE-based neural topic models has focused on enhancing the encoder network by incorporating pre-trained language models (PLMs) and refining topic-word relationships within the generative process. Despite these improvements, the integration of PLMs often results in increased inference costs, and document-topic distributions can still exhibit suboptimal representation. Additionally, existing neural topic models have not addressed the topic-cluster relationships. In this study, we present **TopiCOT** (Neural Topic Model Aligning with Pre-trained Clustering and Optimal Transport), a novel VAE-based topic model designed to overcome these limitations. TopiCOT effectively bridges the gap between the document clustering capabilities of PLMs and the core topic model, avoiding the need for direct PLMs integration. Moreover, we model the correlation between topics and pre-trained clusters through the Optimal Transport (OT) problem, which also enhances document representation and efficiently captures topic associations. Experimental results on popular benchmark datasets demonstrate that our method effectively improves document-topic distributions while preserving a high level of topic coherence comparable to other state-of-the-art baselines. Notably, our approach boosts inference speed by about 600 times compared to UTopic, a leading VAE-based method that leverages pre-trained language models.

1 Introduction

Neural topic models (NTMs) represent the advanced evolution of traditional topic model techniques (Hofmann 1999; Blei, Ng, and Jordan 2003) by leveraging the power of neural networks. These models (Wu et al. 2023b; Han et al. 2023; Zhao et al. 2021; Dieng, Ruiz, and Blei 2020), mostly based on Variational Autoencoders (VAEs) (Kingma and Welling 2013), feature an inference encoder that generates document-topic distributions and a generative decoder that uses the encoder's output along with topic-word proportions to recreate the original texts. By learning richer context embeddings (Dieng, Ruiz, and Blei 2020; Zhao et al. 2021) and refining topic-word relationships, these models enhance topic coherence and document representations (Wu, Nguyen, and Luu 2024).

One of the most effective ways to enhance the performance of VAE-based NTMs is by incorporating pre-trained language models (PLMs) (Devlin et al. 2019; Brown et al. 2020), which excel at capturing linguistic patterns and contextual nuances, into the network architecture (Han et al. 2023; Bianchi, Terragni, and Hovy 2021; Bianchi et al. 2021). Representations generated by PLMs can be used as inputs for the encoder network (Bianchi, Terragni, and Hovy 2021; Bianchi et al. 2021), leading to improved document-topic distributions and greater topic coherence. UTopic (Han et al. 2023) further leverages PLMs not only for the encoder but also for constructing the word set in topic representations. However, such methods often do not directly investigate the impact of PLMs on document representations learned by topic models, leading to suboptimal performance. Recently, some other state-of-the-art methods (Wu et al. 2023b, 2024a) that do not utilize PLMs show even better document-topic distributions, as depicted in Figure 1. Additionally, directly using PLMs in the inference network (encoder) can significantly increase inference costs. In this work, we aim to develop a more effective solution for leveraging PLMs in neural topic models, with a particular focus on enhancing clustering.



(a) 20NG dataset

(b) AGNews dataset

Figure 1: Performance of UTopic (Han et al. 2023), a leading topic model using PLM, compared to other SOTA methods (ECRTM (Wu et al. 2023b) and FASTopic (Wu et al. 2024a)) on the two datasets 20 News Groups (**20NG**) (Lang 1995) and **AGNews** (Zhang, Zhao, and LeCun 2015). Despite using PLM, UTopic shows inferior document-topic distributions, as depicted in Purity and NMI metrics.

On the other hand, some studies focus on modeling topic-word relationships (Dieng, Ruiz, and Blei 2020; Zhao et al. 2021; Wang et al. 2022; Xu et al. 2022, 2023). Most of these approaches model topics within the word embedding space and use various techniques to establish topic-word correlations (Dieng, Ruiz, and Blei 2020; Zhao et al. 2021; Wang et al. 2022; Wu et al. 2023b, 2024a). Recently, ECRTM (Wu et al. 2023b) introduced a novel approach that directly models the relationship between words and topics via optimal transport, effectively addressing the issue of topic collapse. However, existing neural topic models often lack consideration of the relationships between topics through clustering. In reality, documents are usually organized into clusters, and each cluster contains a set of topics. Encoding the relationship between topics through clusters can lead to a deeper understanding of the connections between topics and potentially improve document representations.

This paper presents TopiCOT, an innovative framework that integrates neural topic modeling with pre-trained clustering and optimal transport. TopiCOT leverages the well-pretrained clustering produced by the robust semantic representation capabilities of pre-trained language models to refine the training process of topic distributions in documents and topic embeddings. Firstly, TopiCOT constructs a regularization between document-topic distributions and cluster proportions in documents. This mechanism effectively improves the document representations generated by the topic model, as proved by the enhanced clustering performance.

Secondly, we capture the relationships between topics and pre-trained document clusters by solving a specifically defined Optimal Transport (OT) problem (Peyré and Cuturi 2018) between them. This OT alignment ensures that semantically similar topics are mapped to cluster centers with similar embeddings, resulting in topics that are closely related in meaning being assigned to the same or nearby clusters. Moreover, since documents are represented by the distribution of topics, aligning topics to well-pretrained clusters with OT also contributes to building better document representations. We conclude the contributions of this paper as follows:

- We introduce an innovative VAE-based Topic Model called TopiCOT that regularizes the document-topic distributions by leveraging well-pretrained clustering from a PLM, thereby enhancing the document representations generated by the topic model.
- We propose a novel topic refinement method that captures the relationships between topics and pre-trained document clusters through Optimal Transport, achieving coherent topics and improved document-topic distributions.
- We conduct extensive experiments on benchmark datasets, showing that TopiCOT effectively improves document-topic distributions, achieves coherent topics, and significantly reduces the inference cost caused by PLMs.

2 Related Work

VAE-based Topic Models. Recently, research has shown significant success in applying VAE architecture to topic modeling (Srivastava and Sutton 2017; Dieng, Ruiz, and Blei 2020; Bianchi et al. 2021; Wang et al. 2022; Wu et al. 2023a; Cvejoski, Sánchez, and Ojeda 2023; Wu et al. 2023b, 2024b). Compared to conventional topic models (Hofmann 1999; Blei, Ng, and Jordan 2003), these models improve efficiency and stability in both topic coherence and document representations. Almost all studies on VAE-based topic models follow two main directions: improving document-topic distributions (encoder) and enhancing the reconstruction phase (decoder).

In terms of encoders, integrating pre-trained language model (Devlin et al. 2019; Brown et al. 2020) is one of the most effective approaches. For instance, Bianchi, Terragni, and Hovy (2021) utilizes the contextual document embeddings from SBERT (Reimers and Gurevych 2019) alongside Bag-of-Words (BoW) as input for the encoder network. UTopic (Han et al. 2023) also uses SBERT (Reimers and Gurevych 2019) to generate term weights, which are incorporated into the reconstruction loss to eliminate unnecessary words. Despite their benefits, these approaches often lead to longer inference times. Other approaches do not depend on pre-trained language model and instead use optimal transport distance to measure the differences between documents and topics (Wang et al. 2022; Zhao et al. 2021). By using the document clustering results from PLMs only during the training phase, our method could take advantage of the linguistic features of PLMs without affecting inference time.

Regarding decoders, utilizing pre-trained word embeddings is an effective method for enhancing topic coherence (Dieng, Ruiz, and Blei 2020; Bianchi, Terragni, and Hovy 2021; Wu et al. 2023b). Dieng, Ruiz, and Blei (2020) utilizes word embeddings, such as GloVe (Pennington, Socher, and Manning 2014) and Word2Vec (Mikolov et al. 2013) to create topics with more semantically related terms. In other work, Xu et al. (2022) uses hyperbolic embeddings for topic taxonomy while Xu et al. (2023) improves clusterability by utilizing spherical embeddings. In the embedding space, the relation between topics and words is modeled by utilizing a similarity function (Dieng, Ruiz, and Blei 2020; Nguyen et al. 2022; Zhao et al. 2021) or optimal (conditional) transport distance (Wu et al. 2023b; Wang et al. 2022). To mitigate the topic collapse problem, Wu et al. (2023b) refine each topic embedding as the center of a separate cluster of word embeddings. Furthermore, some approaches aim to construct hierarchical structures of the generated topics (Griffiths et al. 2003; Xu et al. 2022; Wu et al. 2024b). In contrast, our study explores the relationship between topics and pre-trained clusters produced by a pre-trained language model.

Other Approaches in Topic Models. Another line of research in neural topic models is directly generating topics by clustering the document representations (Grootendorst 2022; Sia, Dalmia, and Mielke 2020; Zhang et al. 2022). While this approach is efficient and produces coherent topics, determining the topic proportions for a document is not straightforward. Instead of using the clustering results as the final topics, we use these high-quality clusters to regularize the document-topic distributions and topic embeddings.

Wu et al. (2024a) introduces a new paradigm for topic modeling called FASTopic, which effectively identifies latent topics by reconstructing semantic relationships among documents, topics, and word embeddings, using only optimal transport distance. Additionally, some recent research employs large language models to frame topics as conceptual descriptions (Wang et al. 2023; Pham et al. 2024). However, these methods fall short in providing the proportions of words within topics or topics within documents, which complicates direct comparison with other approaches.

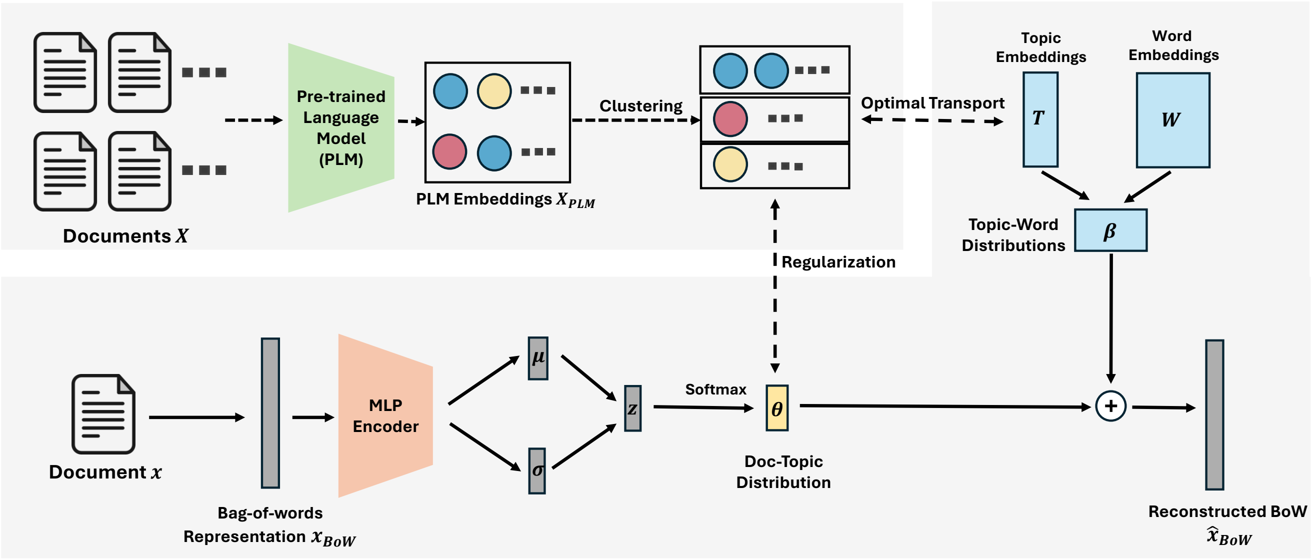


Figure 2: Illustration of our proposed neural topic model, TopiCOT, showing the components excluded during inference time highlighted by the dashed line. TopiCOT leverages high-quality clustering produced by PLM to regularize document-topic distributions and topics construction of the topic model.

3 Background

Notations. Denote $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^D$ be a collection of D documents with the vocabulary of V words. The BoW and pre-trained language model representations of document $\mathbf{x}_i$ are $\mathbf{x}_{i\text{BoW}}$ and $\mathbf{x}_{iPLM}$, respectively. We have $\beta = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^{V \times K}$ as the topic-word distributions of K desired topics. With L as the word embedding dimension, we set $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_V) \in \mathbb{R}^{V \times L}, \mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_K) \in \mathbb{R}^{K \times L}$ be the word embeddings and topic embeddings, respectively. Each document $\mathbf{x}_i$ has the topic proportion $\theta_i \in \mathbb{R}^K$ referring to what topic it includes. $\mathbb{1}_N$ is a vector of length N where every element is 1. $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$ represent the inner product between two same size matrices X and Y. $H(P) = -\langle P, \log P - 1 \rangle = -\sum_{i,j} P_{ij}(\log P_{ij} - 1)$ is the Shannon entropy of P. The KL divergence between P and Q is defined as $\text{KL}(P \| Q) = \sum_{i,j} P_{ij} \log(P_{ij}/Q_{ij})$.

3.1 VAE-based Neural Topic Models

VAE-based neural topic models aim to (i) identify K topics across documents in the corpus $\mathbf{X}$ and (ii) infer topic proportions for a new document x. For the first objective, we represent the topics as a matrix $\beta \in \mathbb{R}^{K \times V}$, where each row $\beta_i \in \mathbb{R}^V$ corresponds to a multinomial probability distribution over the vocabulary of a topic. Commonly, the matrix β is decomposed as the product of word embeddings $\mathbf{W}$ and topic embeddings $\mathbf{T}$ (Dieng, Ruiz, and Blei 2020; Xu et al. 2022). However, Wu et al. (2023b) define β as follows:

$$\beta_{ij} = \frac{\exp\left(-\|\mathbf{w}_i - \mathbf{t}_j\|^2/\tau\right)}{\sum_{j'=1}^K \exp\left(-\|\mathbf{w}_i - \mathbf{t}_{j'}\|^2/\tau\right)}$$

where τ is a temperature hyperparameter. The word embeddings $\mathbf{W}$ are typically initialized using pre-trained embeddings such as GloVe (Pennington, Socher, and Manning 2014) or Word2Vec (Mikolov et al. 2013).

Regarding the second objective of Neural Topic Models, the topic proportion θ is modeled to depend on a latent variable z, which conforms to a logistic-normal distribution characterized as $p(z) = \mathcal{LN}(z|\mu_0, \Sigma_0)$. A document $\mathbf{x}$ is first represented as its BoW representation $\mathbf{x}_{\text{BoW}}$, then encoded through neural networks to obtain the parameters of a normal distribution, with mean $\mu = h_\mu(\mathbf{x}_{\text{BoW}})$ and diagonal covariance matrix $\Sigma = \text{diag}(h_\Sigma(\mathbf{x}_{\text{BoW}}))$. Subsequently, z is sampled from the posterior distribution $q(z|\mathbf{x}) = \mathcal{N}(z|\mu, \Sigma)$ using the reparameterization trick (Kingma and Welling 2013). The softmax function is then applied to z, yielding the topic proportion $\theta = \text{softmax}(z)$. The BoW representation is reconstructed with the topic-word distribution matrix β from a multinomial distribution $\hat{\mathbf{x}}_{\text{BoW}} \sim \text{Multi}(\text{softmax}(\beta\theta))$. The topic modeling loss is constructed to include a reconstruction term and a regularization term as follows:

$$\mathcal{L}_{\text{TM}} = \frac{1}{D} \sum_{i=1}^{D} \Big[-(\mathbf{x}_{i\text{BoW}})^\top \log(\text{softmax}(\beta\theta_i)) + \text{KL}(q(z|\mathbf{x}_i) \| p(z)) \Big].$$

3.2 Entropic Regularized Optimal Transport

Let A and B be two discrete measures on the supports of n and m points, respectively, associate with the weights $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_m)$ satisfying $\sum_i a_i = \sum_j b_j$. Given a cost matrix $C \in \mathbb{R}^{n \times m}$, an optimal transport plan P of an entropic regularized optimal transport problem is defined as the solution to the following optimization

Algorithm 1: Learning TopiCOT

```
Require: Document collection $\mathbf{X}$, pre-trained language model $\mathrm{PLM}$, pre-trained word embedding $\mathbf{W}_{\mathrm{GloVe}}$, number of topic $K$;
Ensure: Encoder network's parameter $W_{\mathrm{enc}}$, linear projections' parameter $W_{\phi_\theta}$ and $W_{\phi_T}$, word embedding $\mathbf{W}$, topic embedding $\mathbf{T}$, topic-cluster transport plan $\pi^*$;
  Initialize $\mathbf{W} = \mathbf{W}_{GloVe}$;
  Calculate embedding of the documents $\mathbf{X}_{\mathrm{PLM}}$;
  Clustering the documents embedding using KMeans;
  Calculate the distance matrix $P$ between documents PLM embeddings and cluster embeddings;
  repeat
    Calculate the distance matrix $Q$ between linear transformed doc-topic distribution and cluster embeddings;
    Update $W_{\phi_\theta}$ using a gradient descent step;
    Calculate $\pi^*$ using Sinkhorn algorithm;
    Update $W_{\phi_T}$ using a gradient descent step;
    Update $W_{\mathrm{enc}}, \mathbf{T}, \mathbf{W}$ using a gradient descent step;
  until converge
```

problem (Peyré and Cuturi 2018):

$$\begin{aligned} &\text{minimize } \langle P, C \rangle - \epsilon H(P) \\ &\text{s.t. } P \in \mathbb{R}^{n \times m} \\ &P\mathbb{1} = A, P^\top \mathbb{1} = B \end{aligned} \tag{1}$$

The entropic regularization terms allow us to use iterative algorithms such as the Sinkhorn algorithm (Cuturi 2013) to solve this optimization problem efficiently.

4 Proposed Method

We enhance both the inference network (encoder) and the generative model (decoder) of neural topic models through pre-trained clustering regularization. The model details are depicted in Figure 2.

4.1 Clustering Regularization on Doc-Topic Distribution

We introduce a new regularization for doc-topic distribution based on Jeffreys divergence, which can enhance the inference network. We use a pre-trained language model to generate embedding for all documents: $X_{PLM} \in \mathbb{R}^{D \times M}$ where M is the document embedding size. After that, we employ the K-means clustering method to partition the D documents into G clusters. We denote the cluster centers set as $(E_1, E_2, ..., E_G)$ where each $E_i \in \mathbb{R}^M$.

We form a matrix $P \in \mathbb{R}^{D \times G}$ that demonstrates the cluster proportions of documents that:

$$P_{ij} = \frac{d_{ij}}{\sum_{g=1}^{G} d_{ig}} \tag{2}$$

where d_{ij} is the distance between document i and the center of cluster j. We hope that the topic distribution of documents can achieve as good clustering as the document embeddings generated from the pre-trained language model. To attain this, we project the doc-topic distribution onto the document embedding space: $\phi_\theta(\theta)$, where ϕ_θ is a learnable linear projection with learnable weights $W_{\phi_\theta} \in \mathbb{R}^{K \times M}$. A matrix $Q \in \mathbb{R}^{D \times G}$ is formed similarly to matrix P, but d_{ij} in Q is the distance between the projection of the doc-topic distribution and the cluster's center. Then we try to minimize the following Jeffreys divergence between P and Q, we call this Document Clustering Regularization (DCR):

$$\mathcal{L}_{\mathrm{DCR}} = \mathrm{KL}\left(P \| Q\right) + \mathrm{KL}\left(Q \| P\right) \tag{3}$$

4.2 Clustering Regularization on Topic Embedding

Besides doc-topic distribution, we introduce another regularization on topic embedding which is taken from optimal transport with cluster embedding. Given that documents are assumed to be delivered in groups with similar semantic meanings, it is assumed that the topics also exhibit a cluster structure. Given that the size of the cluster g is n_g and the cluster size proportion as $s_g = n_g/D$, we have $s = (s_1, ..., s_G)$ as the vector of all cluster size proportions.

We formulate the relationship of topic embeddings and document clusters with the transport plan of a specifically defined optimal transport problem. Particularly, we define two discrete measures of topics and clusters as: $\alpha = \sum_{k=1}^{K} \frac{1}{K} \delta_{\mathbf{t}_k}$ and $\beta = \sum_{g=1}^{G} s_g \delta_{E_g}$, where δ_x denotes the Dirac unit mass on x. The transportation cost between a topic k and a cluster g is denoted by: $C_{TE} = \|\phi_T(\mathbf{t}_k) - E_g\|^2$, where ϕ_T is a learnable linear projection from space of topic embedding to cluster embedding, with learnable weights $W_{\phi_T} \in \mathbb{R}^{L \times M}$. The optimal transport plan π^* is the solution for the following optimization problem:

$$\begin{aligned} &\text{minimize } \langle C_{\mathrm{TE}}, \pi \rangle - \nu H(\pi) \\ &\text{s.t. } \pi \in \mathbb{R}^{T \times G} \\ &\pi \mathbb{1}_G = \frac{1}{K} \mathbb{1}_K, \pi^T \mathbb{1}_K = s \end{aligned} \tag{4}$$

Subsequently, the Sinkhorn algorithm is employed to solve the optimization problem (Cuturi 2013). By doing so, π^* is a differentiable variable parameterized by transport cost matrix C_{TE} (Salimans et al. 2018; Genevay, Peyre, and Cuturi 2018). After that, we employ the Topic Clustering Regularization (TCR) objective by minimizing the total distance between topic and cluster embeddings weighted by transport plan:

$$\mathcal{L}_{\mathrm{TCR}} = \sum_{k=1}^{K} \sum_{g=1}^{G} \|\phi_T(\mathbf{t}_k) - E_g\|^2 \pi^*_{kg} \tag{5}$$

As π^* is a solution of an optimal transport problem, it inherits the sparsity property of a transport plan. In other words, only a few transport pairs π^*_{kg} are significant. When minimizing $\mathcal{L}_{\mathrm{TCR}}$, the distance between corresponding projected topic embeddings and cluster embeddings is minimized. This results in a cluster structure where the cluster embeddings are the centers of the clusters of projected topic embeddings.

Model	20NG				IMDB				AGNews			
	Purity	NMI	$\mathrm{C_v}$	TD	Purity	NMI	$\mathrm{C_v}$	TD	Purity	NMI	$\mathrm{C_v}$	TD
LDA ‡	0.367	0.364	0.385	0.655	0.614	0.041	0.347	0.788	0.640	0.193	0.364	0.864
ETM ‡	0.347	0.319	0.375	0.704	0.660	0.038	0.346	0.557	0.679	0.224	0.364	0.819
NSTM ‡	0.354	0.356	0.395	0.427	0.658	0.040	0.334	0.175	0.772	0.324	0.411	0.877
WeTe ‡	0.268	0.304	0.383	0.949	0.587	0.031	0.368	0.931	0.641	0.268	0.383	0.945
ECRTM ‡	0.560	0.524	0.431	0.964	0.694	0.058	0.393	**0.974**	0.802	0.367	0.466	0.961
UTopic	0.530	0.454	**0.508**	0.860	0.550	0.005	**0.429**	0.554	0.768	0.303	**0.545**	0.838
FASTopic	0.583	0.528	0.427	**0.980**	0.683	0.055	0.371	0.969	0.831	0.352	0.379	0.960
TopiCOT	**0.638**	**0.573**	0.449	0.863	**0.706**	**0.059**	0.392	0.925	**0.839**	**0.405**	0.454	**0.999**

Table 1: Performance evaluation across three datasets for 50 topics, assessed by mean Purity, mean NMI, mean $\mathrm{C_v}$, and mean TD. The best results are highlighted in **bold** and the second-best are underlined. ‡ Results reported in (Wu et al. 2023b).

Model	20NG				IMDB				AGNews			
	Purity	NMI	$\mathrm{C_v}$	TD	Purity	NMI	$\mathrm{C_v}$	TD	Purity	NMI	$\mathrm{C_v}$	TD
LDA ‡	0.364	0.346	0.387	0.622	0.600	0.037	0.342	0.691	0.654	0.194	0.349	0.696
ETM ‡	0.394	0.339	0.369	0.573	0.648	0.037	0.341	0.371	0.674	0.204	0.371	0.773
NSTM ‡	0.383	0.363	0.391	0.473	0.659	0.039	0.340	0.255	0.764	0.359	0.421	0.832
WeTe ‡	0.338	0.348	0.352	0.742	0.589	0.025	0.293	0.638	0.699	0.271	0.363	0.827
ECRTM ‡	0.555	0.494	0.405	**0.904**	0.694	0.049	0.373	**0.887**	0.812	0.428	0.416	**0.981**
UTopic	0.545	0.452	**0.523**	0.750	0.553	0.004	**0.534**	0.656	0.760	0.283	**0.548**	0.681
FASTopic	**0.622**	0.522	0.400	0.861	0.680	0.048	0.369	0.886	0.833	0.330	0.385	0.912
TopiCOT	**0.622**	**0.523**	0.407	0.873	**0.711**	**0.055**	0.378	0.797	**0.839**	**0.449**	0.422	0.856

Table 2: Performance evaluation across three datasets for 100 topics, assessed by mean Purity, mean NMI, mean $\mathrm{C_v}$, and mean TD. The best results are highlighted in **bold** and the second-best are underlined. ‡ Results reported in (Wu et al. 2023b).

4.3 Overall Objective Function

Our inference process and topic modeling loss function follow the conventional neural topic model, as noted in Section 3. Additionally, inspired by (Wu et al. 2023b), we employ the Embedding Clustering Regularization regularizer to mitigate the topic collapsing problem:

$$\mathcal{L}_{\mathrm{ECR}} = \sum_{i=1}^{V}\sum_{j=1}^{K} \|\mathbf{w}_i - \mathbf{t}_j\|^2 \epsilon^*_{ij} \tag{6}$$

where ϵ^* is the solution of the following optimization problem:

$$\begin{aligned} &\text{minimize } \langle C_{\mathrm{WT}}, \epsilon\rangle - \nu H(\epsilon) \\ &\quad \text{s.t. } \epsilon \in \mathbb{R}^{V\times K} \\ &\quad\quad \epsilon \mathbb{1}_K = \frac{1}{V}\mathbb{1}_V, \epsilon^T \mathbb{1}_V = \frac{1}{K}\mathbb{1}_K \end{aligned} \tag{7}$$

where $C_{\mathrm{WT}} \in \mathbb{R}^{V\times K}$ is the distance matrix between word embeddings and topic embeddings. ϵ^* is obtained using the Sinkhorn algorithm (Cuturi 2013).

In summary, our overall objective function is described below:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{TM}} &+ \lambda_{\mathrm{ECR}}\mathcal{L}_{\mathrm{ECR}} \\ &+ \lambda_{\mathrm{DCR}}\mathcal{L}_{\mathrm{DCR}} + \lambda_{\mathrm{TCR}}\mathcal{L}_{\mathrm{TCR}} \end{aligned} \tag{8}$$

where $\lambda_{\mathrm{ECR}}, \lambda_{\mathrm{DCR}}, \lambda_{\mathrm{TCR}}$ are weight hyperparameters. The detailed training algorithm for TopiCOT is presented in Algorithm 1.

5 Experiments

5.1 Experimental Settings

Datasets. We used four popular datasets for topic modeling, including 20 News Groups (**20NG**) (Lang 1995), **AGNews** (Zhang, Zhao, and LeCun 2015), **IMDB** (Maas et al. 2011), and **BBC** (Greene and Cunningham 2006).

Pre-trained Cluster. We utilized the all-MiniLM-L6-v2 model (Reimers and Gurevych 2019) for pre-trained language model. Clustering was conducted to optimize the quality for each dataset, resulting in ideal numbers of PLM clusters, which should be smaller than the number of topics: 20 for 20NG, 4 for IMDB, 3 for AGNews and 4 for BBC.

Evaluation Metrics. We followed the evaluation approaches proposed by (Wu et al. 2023b) to assess both document-topic distribution quality and topic quality, with a focus on the former. The quality of document-topic distributions is evaluated using NMI and Purity metrics (Manning, Raghavan, and Schütze 2008) when using doc-topic proportions in the document clustering task. Topic quality is assessed through metrics of topic coherence and topic diversity. We follow (Röder, Both, and Hinneburg 2015) to measure topic coherence with $\mathrm{C_v}$, NPMI, and $\mathrm{C_p}$, the metrics having strong correlation with human judgment. These coherence metrics are computed using a version of the Wikipedia corpus[1] as an external reference. Additionally, NPMI is calculated using the training dataset as a reference, denoted as $\mathrm{NPMI} - \mathrm{I}$, alongside

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

Model	20NG						BBC					
	10 topics			*20 topics*			*10 topics*			*20 topics*		
	NPMI-W	NPMI-I	C_p	NPMI-W	NPMI-I	C_p	NPMI-W	NPMI-I	C_p	NPMI-W	NPMI-I	C_p
LDA †	-0.0056	0.0661	0.0719	0.0057	0.0801	0.0727	-0.0718	-0.0205	-0.0709	-0.0746	-0.0199	-0.0684
ProdLDA †	-0.0227	-0.0083	-0.0634	-0.0158	-0.0610	-0.0429	0.0084	0.0110	0.0569	0.0001	-0.0105	0.0647
ETM †	0.0234	0.0927	0.1207	0.0052	0.1219	0.0527	-0.0333	0.0251	0.0416	-0.0212	0.0441	0.0829
CTM †	-0.0086	0.1149	0.0156	-0.0161	0.1244	-0.1415	0.0289	0.1109	0.3254	0.0436	0.0714	0.2543
ClusterTM †	0.0154	-0.2863	0.0082	0.0135	-0.2870	0.0160	0.0339	0.0990	0.0908	0.0255	0.0656	0.0588
BertTopic †	0.0322	-0.0563	0.1515	0.0609	-0.0903	0.2318	0.0456	0.0762	0.2556	-0.0007	0.0943	0.0747
UTopic †	0.0653	0.1231	0.3709	**0.1069**	0.1130	**0.4850**	0.0708	0.1018	0.3925	0.0938	0.1256	**0.5388**
TopiCOT	**0.0803**	**0.1600**	**0.4205**	0.0880	**0.1521**	0.4358	**0.0838**	**0.1560**	**0.4175**	**0.0973**	**0.1766**	0.4253

Table 3: Topic coherence comparison on the **20NG** and **BBC** datasets for models with 10 and 20 topics. The best results are highlighted in **bold** and the second-best are underlined. †Results reported in (Han et al. 2023).

Model	20NG				AGNews			
	Purity	NMI	C_v	TD	Purity	NMI	C_v	TD
ECRTM ‡	0.560	0.524	0.431	**0.964**	0.802	0.367	**0.466**	0.961
TopiCOT	**0.638**	**0.573**	**0.449**	0.863	**0.839**	**0.405**	0.454	**0.999**
w/o TCR	0.596	0.549	0.433	0.921	0.824	0.376	0.462	0.868
w/o DCR	0.616	0.558	0.434	0.895	0.823	0.361	0.459	0.984

Table 4: Ablation study on **20NG** and **AGNews** datasets for models with 50 topics. The best results are highlighted in **bold** and the second-best are underlined. ‡ Results reported in (Wu et al. 2023b).

the version based on the Wiki corpus ($\mathrm{NPMI-W}$). Topic diversity (TD) is quantified by the proportion of unique words among the topic words.

Baseline Models. We consider the following baseline models for comparison: **LDA** (Blei, Ng, and Jordan 2003), a well-established probabilistic topic model; **ProdLDA** (Srivastava and Sutton 2017), a LDA-variant incorporating Variational Autoencoders (VAEs); **ETM** (Dieng, Ruiz, and Blei 2020), a neural topic model integrating word embeddings; **NSTM** (Zhao et al. 2021) and **WeTe** (Wang et al. 2022), the NTMs which use optimal and conditional transport distance to model reconstruction loss, respectively; **ECRTM** (Wu et al. 2023b), a NTM implements word-topic transport plan to prevent topic collapse problem; **CTM** (Bianchi, Terragni, and Hovy 2021) and **UTopic** (Han et al. 2023), the NTMs incorporate PLM's contextualized embedding as input of the encoder; **FASTopic** (Wu et al. 2024a), which employs a regularization based on modeling the semantic relationship between topics and words as an OT plan; **ClusterTM** (Sia, Dalmia, and Mielke 2020) and **BERTopic** (Grootendorst 2022), the NTMs which cluster documents based on their contextual embeddings and use term frequency (tf) to generate topic words.

5.2 Doc-Topic Distribution and Topic Quality

We initially conducted experiments to evaluate the effectiveness of our method by assessing the quality of document-topic distributions (Purity and NMI) and general topic quality (C_v and TD). We utilized three datasets: 20NG, IMDB and AGNews, with pre-processing steps derived from ECRTM (Wu et al. 2023b). In addition to traditional models like LDA and ETM, we compared our approach against recent leading methods in topic modeling. Tables 1 and 2 show the performance of these methods for 50 and 100 topics, respectively. UTopic, which mostly focuses on enhancing topic coherence, excels in this metric but under-performs in document-topic distribution quality. In contrast, our method, which incorporates PLM clustering performance through both document and topic regularization, demonstrates superior Purity and NMI across all datasets. Regarding topic coherence, TopiCOT ranks as the second-best method after UTopic and shows improvements over ECRTM backbone models, thanks to its well-trained clustering regularization. For topic diversity, models employing clustering regularization on topic and word embeddings (ECRTM, FASTopic, TopiCOT) exhibit significantly higher results. Although TopiCOT's TCR loss leads to a soft-assignment of topics to clusters and brings topics closer, reducing diversity, this reduction is not substantial. TopiCOT's TD score remains competitive, just behind ECRTM and comparable to FASTopic.

To further validate the topic quality of TopiCOT, we replicated the experimental settings of UTopic (Han et al. 2023) to compare our method with PLM-based approaches, especially clustering-based topic models which only concentrate on constructing coherent topics and struggle with assigning a mixture of topics to each document. We used the 20NG and BBC datasets with pre-processing procedures outlined in (Han et al. 2023), the ideal number of clusters in 20NG that is smaller than the number of topics is chosen to be four. The results, presented in Tables 3, show that our approach demonstrates higher performances compared to UTopic and consistently outperforms other methods. By leveraging pre-trained language model clustering to build relationships between documents and topic representations, TopiCOT enhances the model's ability to capture topic relationships and effectively improves topic quality.

5.3 Ablation Study

We conducted ablation experiments on 20NG and AGNews to assess the impact of each component on our method's overall performance. Specifically, we removed both Topic Cluster-

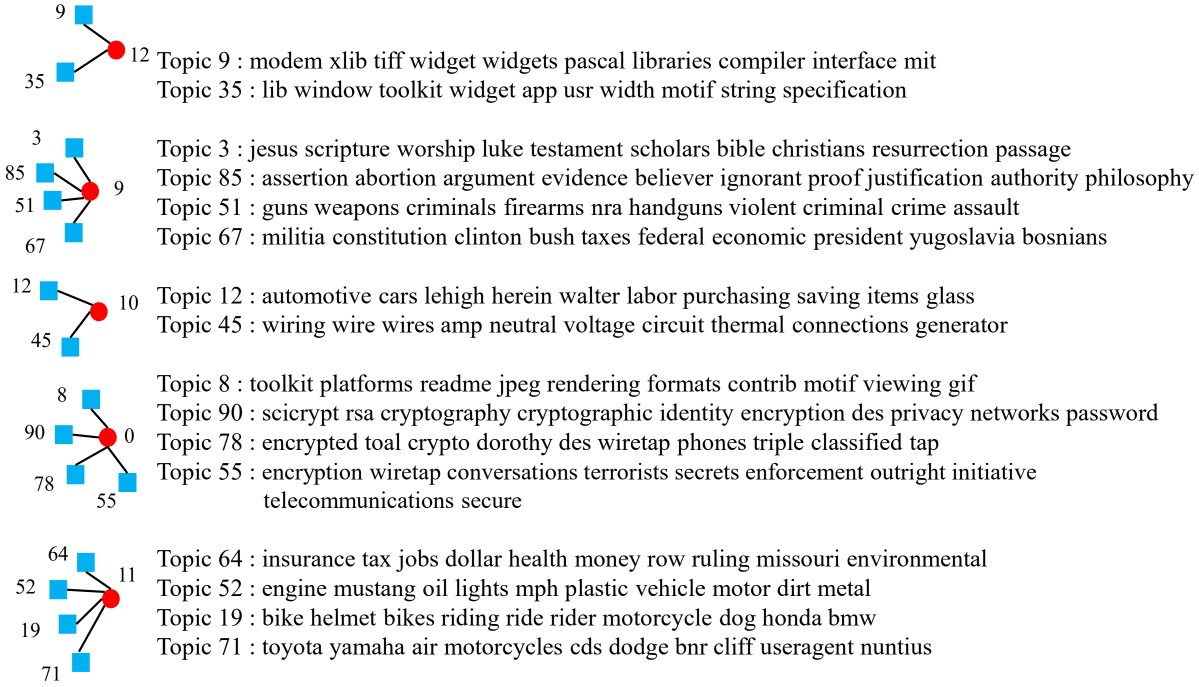

9
35
12
Topic 9 : modem xlib tiff widget widgets pascal libraries compiler interface mit
Topic 35 : lib window toolkit widget app usr width motif string specification
3
85
51
67
9
Topic 3 : jesus scripture worship luke testament scholars bible christians resurrection passage
Topic 85 : assertion abortion argument evidence believer ignorant proof justification authority philosophy
Topic 51 : guns weapons criminals firearms nra handguns violent criminal crime assault
Topic 67 : militia constitution clinton bush taxes federal economic president yugoslavia bosnians
12
45
10
Topic 12 : automotive cars lehigh herein walter labor purchasing saving items glass
Topic 45 : wiring wire wires amp neutral voltage circuit thermal connections generator
8
90
78
55
0
Topic 8 : toolkit platforms readme jpeg rendering formats contrib motif viewing gif
Topic 90 : scicrypt rsa cryptography cryptographic identity encryption des privacy networks password
Topic 78 : encrypted toal crypto dorothy des wiretap phones triple classified tap
Topic 55 : encryption wiretap conversations terrorists secrets enforcement outright initiative telecommunications secure
64
52
19
71
11
Topic 64 : insurance tax jobs dollar health money row ruling missouri environmental
Topic 52 : engine mustang oil lights mph plastic vehicle motor dirt metal
Topic 19 : bike helmet bikes riding ride rider motorcycle dog honda bmw
Topic 71 : toyota yamaha air motorcycles cds dodge bnr cliff useragent nuntius

Figure 3: Visualization of the transport plan illustrating the connections between selected topics (blue) and clusters (red) learned from the 20 News Groups (20NG) dataset. The lines represent significant connections between topics i and clusters j, as determined by the transport proportion π^*_{ij}, with a threshold of 9.9×10^{-3}.

Model	**20NG**	**IMDB**	**AGNews**
UTopic	44.456	98.689	39.475
TopiCOT	0.074	0.145	0.066
Speedup	603.475	680.614	595.101

Table 5: Comparison of inference times (in seconds) on an NVIDIA RTX 3060 GPU for UTopic and TopiCOT models across three datasets with 50 topics.

ing Regularization and Document Clustering Regularization individually and evaluated the resulting model performance. Table 4 presents the performance results, including ECRTM, a version of TopiCOT that excludes all proposed techniques. Our findings show that both components enhance the model's performance in terms of document-topic distribution quality, as measured by Purity and NMI metrics. Although integrating pre-trained clustering regularization slightly reduces topic diversity, it adequately achieves coherent topics. This effect is expected, as the pre-trained clusters help capture topic relationships but also bring topics closer together, which may reduce their differences and overall diversity.

5.4 Inference Time

We conducted experiments to compare the inference time of our model with UTopic (Han et al. 2023), a leading VAE-based topic model that utilizes PLM. We assess the total time taken to pass the entire training set through the encoder network, using a batch size of 128. As shown in Table 5, TopiCOT reduces inference costs by approximately 600 times compared to UTopic. Additionally, it delivers significantly better document-topic distribution quality and comparable topic quality, as detailed in Subsection 5.2. These results validate the effectiveness of our proposed methods in utilizing PLM for topic modeling while overcoming its limitations associated with high inference costs.

5.5 Topic-Cluster Relationship Visualization

To elucidate the structure of the learned transport plan, we present in Figure 3 a visualization of five clusters, along with their associated topics and corresponding topic words. We highlight the significant topic-cluster pairs. Our observations reveal that topics within the same cluster tend to exhibit semantically similar words, demonstrating the method's capability to group topics into semantically coherent clusters.

6 Conclusion

In this study, we introduce **TopiCOT**, an innovative framework aimed at enhancing the performance of VAE-based neural topic modeling. TopiCOT refines the document-topic distributions by the regularization with the text clustering capabilities of Pre-trained Language Model and utilizes Optimal Transport to capture the relationships between topics and pre-trained clusters. The experimental results present the effectiveness of TopiCOT in enhancing the document-topic distributions, achieving coherent topics and significantly reducing inference time.

References

Bianchi, F.; Terragni, S.; and Hovy, D. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *ACL-IJCNLP (Volume 2: Short Papers)*, 759–766. Association for Computational Linguistics.

Bianchi, F.; Terragni, S.; Hovy, D.; Nozza, D.; and Fersini, E. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. Association for Computational Linguistics.

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

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, 26. Curran Associates, Inc.

Cvejoski, K.; Sánchez, R. J.; and Ojeda, C. 2023. Neural dynamic focused topic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12719–12727.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 4171–4186. Association for Computational Linguistics.

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

Genevay, A.; Peyre, G.; and Cuturi, M. 2018. Learning Generative Models with Sinkhorn Divergences. In Storkey, A.; and Perez-Cruz, F., eds., *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 1608–1617. PMLR.

Greene, D.; and Cunningham, P. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, 377–384. Association for Computing Machinery.

Griffiths, T.; Jordan, M.; Tenenbaum, J.; and Blei, D. 2003. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 16.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.

Han, S.; Shin, M.; Park, S.; Jung, C.; and Cha, M. 2023. Unified Neural Topic Model via Contrastive Learning and Term Weighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1802–1817. Association for Computational Linguistics.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Lang, K. 1995. NewsWeeder: Learning to Filter Netnews. In *Machine Learning Proceedings 1995*, 331–339. San Francisco (CA): Morgan Kaufmann.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Association for Computational Linguistics.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press. ISBN 0521865719.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*.

Nguyen, T.; Mai, T.; Nguyen, N.; Van, L. N.; and Than, K. 2022. Balancing stability and plasticity when learning topic models from short and noisy text streams. *Neurocomputing*, 505: 30–43.

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

Peyré, G.; and Cuturi, M. 2018. Computational Optimal Transport. *Found. Trends Mach. Learn.*, 11(5-6): 355–607.

Pham, C. M.; Hoyle, A.; Sun, S.; and Iyyer, M. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2956–2984.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Association for Computational Linguistics.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. Association for Computing Machinery.

Salimans, T.; Zhang, H.; Radford, A.; and Metaxas, D. N. 2018. Improving GANs Using Optimal Transport. In *ICLR (Poster)*.

Sia, S.; Dalmia, A.; and Mielke, S. J. 2020. 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)*, 1728–1736. Association for Computational Linguistics.

Srivastava, A.; and Sutton, C. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Wang, D.; Guo, D.; Zhao, H.; Zheng, H.; Tanwisuth, K.; Chen, B.; and Zhou, M. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

Wang, H.; Prakash, N.; Hoang, N.; Hee, M.; Naseem, U.; and Lee, R. 2023. Prompting Large Language Models for Topic Modeling. In *2023 IEEE International Conference on Big Data (BigData)*, 1236–1241. Los Alamitos, CA, USA: IEEE Computer Society.

Wu, X.; Dong, X.; Nguyen, T.; Liu, C.; Pan, L.-M.; and Luu, A. T. 2023a. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13763–13771.

Wu, X.; Dong, X.; Nguyen, T. T.; and Luu, A. T. 2023b. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, 37335–37357.

Wu, X.; Nguyen, T.; and Luu, A. T. 2024. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2): 1–30.

Wu, X.; Nguyen, T.; Zhang, D. C.; Wang, W. Y.; and Luu, A. T. 2024a. FASTopic: A Fast, Adaptive, Stable, and Transferable Topic Modeling Paradigm. arXiv:2405.17978.

Wu, X.; Pan, F.; Nguyen, T.; Feng, Y.; Liu, C.; Nguyen, C.-D.; and Luu, A. T. 2024b. On the Affinity, Rationality, and Diversity of Hierarchical Topic Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19261–19269.

Xu, W.; Jiang, X.; Sengamedu Hanumantha Rao, S.; Iannacci, F.; and Zhao, J. 2023. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4433–4457. Association for Computational Linguistics.

Xu, Y.; Wang, D.; Chen, B.; Lu, R.; Duan, Z.; Zhou, M.; et al. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In *Advances in Neural Information Processing Systems*, 35, 31557–31570. Curran Associates, Inc.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, 28. Curran Associates, Inc.

Zhang, Z.; Fang, M.; Chen, L.; and Namazi Rad, M. R. 2022. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3886–3893. Association for Computational Linguistics.

Zhao, H.; Phung, D.; Huynh, V.; Le, T.; and Buntine, W. 2021. Neural Topic Model via Optimal Transport. In *9th International Conference on Learning Representations, ICLR 2021*.

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (**yes**/partial/no/NA)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (**yes**/no)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (**yes**/no)

Does this paper make theoretical contributions? (yes/**no**)
Does this paper rely on one or more datasets? (**yes**/no)

- A motivation is given for why the experiments are conducted on the selected datasets (**yes**/partial/no/NA)
- All novel datasets introduced in this paper are included in a data appendix. (yes/partial/no/**NA**). We do not introduce novel datasets in this study.
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/-partial/no/**NA**). We do not introduce novel datasets in this study.
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (**yes**/no/NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (**yes**/partial/no/NA)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes/partial/no/**NA**). We do not use datasets that are not publicly available.

Does this paper include computational experiments? (**yes**/no)

- Any code required for pre-processing data is included in the appendix. (**yes**/partial/no).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (**yes**/partial/no)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (**yes**/partial/no)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (**yes**/partial/no)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (**yes**/partial/no/NA)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (**yes**/partial/no)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (**yes**/partial/no)
- This paper states the number of algorithm runs used to compute each reported result. (**yes**/no)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (**yes**/no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/**no**)
- This paper lists all final (hyper-)parameters used for each model or algorithm in the paper's experiments. (**yes**/partial/no/NA)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (**yes**/partial/no/NA)