NeuroMax: Enhancing Neural Topic Modeling via Maximizing Mutual Information and Group Topic Regularization

Duy-Tung Pham[1, 2*], **Thien Trang Nguyen Vu**[3*], **Tung Nguyen**[1*], **Linh Van Ngo**[1†],
Duc Anh Nguyen[1], **Thien Huu Nguyen**[4],
[1] Hanoi University of Science and Technology, Vietnam, [2] FPT Software AI Center, Vietnam,
[3] VinAI Research, Vietnam, [4] University of Oregon, USA,

Abstract

Recent advances in neural topic models have concentrated on two primary directions: the integration of the inference network (encoder) with a pre-trained language model (PLM) and the modeling of the relationship between words and topics in the generative model (decoder). However, the use of large PLMs significantly increases inference costs, making them less practical for situations requiring low inference times. Furthermore, it is crucial to simultaneously model the relationships between topics and words as well as the interrelationships among topics themselves. In this work, we propose a novel framework called NeuroMax (**Neur**al T**o**pic Model with **Max**imizing Mutual Information with Pretrained Language Model and Group Topic Regularization) to address these challenges. NeuroMax maximizes the mutual information between the topic representation obtained from the encoder in neural topic models and the representation derived from the PLM. Additionally, NeuroMax employs optimal transport to learn the relationships between topics by analyzing how information is transported among them. Experimental results indicate that NeuroMax reduces inference time, generates more coherent topics and topic groups, and produces more representative document embeddings, thereby enhancing performance on downstream tasks.

1 Introduction

Topic modeling (Hofmann, 1999; Blei et al., 2003; Blei and Lafferty, 2006; Li et al., 2015; Srivastava and Sutton, 2017; Bach et al., 2023; Zhao et al., 2020) is a well-established task in natural language processing (NLP) that involves uncovering and extracting latent topics from extensive corpora, thereby facilitating the comprehension and organization of unstructured data (Kherwa and Bansal, 2019). Its diverse applications span across fields such as text mining (Van Linh et al., 2017; Valero et al., 2022), bioinformatics (Juan et al., 2020), and recommender systems (Le et al., 2018) and streaming learning (Ha et al., 2019; Nguyen et al., 2021; Tuan et al., 2020).

Neural topic models (Srivastava and Sutton, 2017; Wang et al., 2022; Wu et al., 2023b, 2024b; Zhao et al., 2020; Dieng et al., 2020) extend traditional topic modeling methods by incorporating neural network structures, thereby enhancing scalability and efficiency. Similar to Variational Autoencoders (VAEs) (Kingma and Welling, 2013), neural topic models typically consist of two main components: an encoder (inference network) and a decoder (generative network). Recent research has focused on improving these components, leading to overall advancements in model performance.

Regarding the encoder, several studies have proposed incorporating knowledge from pretrained language models (Han et al., 2023; Bianchi et al., 2021b) such as BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018). These models, trained on vast amounts of text data, effectively capture linguistic patterns and contextual information. This rich information can serve as input for the encoder (Bianchi et al., 2021a; Han et al., 2023), enhancing the topic models' ability to generate coherent topics. However, despite such an advantage, utilizing large pretrained models significantly increases inference costs, which limits their utility in scenarios requiring low inference time.

Concerning the decoder, a line of work has leveraged pretrained word embeddings to better capture the semantics of the vocabulary (Dieng et al., 2020; Zhao et al., 2020; Xu et al., 2023, 2022; Nguyen et al., 2022a). Recently, (Wu et al., 2023b) decomposed the topic-word distribution matrix into word and topic embeddings, with the word embeddings initialized by pretrained knowledge. The topic-word distribution is then modeled as the soft-

[*] Equally contributed.

[†] Corresponding author: linhnv@soict.hust.edu.vn

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7758–7772
November 12-16, 2024 ©2024 Association for Computational Linguistics

max of the negative L_2 distance between corresponding embeddings. Additionally, their method employs clustering regularization to group word embeddings into clusters, with each cluster corresponding to a topic, thereby mitigating the topic collapse problem. While these approaches improve efficiency in modeling word-topic relationships, they lack adequate consideration for capturing semantic connections between topics, resulting in a topic embedding space that is difficult to interpret.

In response to these shortcomings, we propose a neural topic model framework that leverages the power of pretrained knowledge without incurring expensive inference costs and effectively captures semantic interrelationships at the topic level. First, for the encoder, we hypothesize that the topic proportions and embeddings derived from pretrained language models should exhibit similar representational characteristics. By maximizing the mutual information between these two variables, we integrate contextualized information from pretrained language models during training, thereby eliminating the need for pretrained components at the inference stage. Figure 1 presents the overall architecture of our proposed encoder, which operates without pretrained language models during inference. Second, inspired by (Van Assel et al., 2023), we employ optimal transport (OT) to model the relationships between topics. Specifically, in line with standard works (Wu et al., 2023b), we assume that each topic carries an equal amount of information, which is transported from one topic to another in a manner that preserves the total information within each topic. The learned transport plan elucidates the connections between the topics. Additionally, we assume that documents often encompass several closely related themes, naturally grouping topics into semantically related clusters. We enhance the group relationship of topics by imposing a regularization on the aforementioned transport plan based on predefined topic cluster relationships derived from topic clustering.

The rest of this paper is organized as follows: Section 2 lists some related works in topic modeling and neural topic modeling. Some background on neural topic models, mutual information maximization, and optimal transport is provided in 3. Our proposed methodologies are introduced in 4. Some experiments to illustrate the effectiveness of the proposed method are reported in 5. Finally, our discussion and conclusion are given in Section 6.

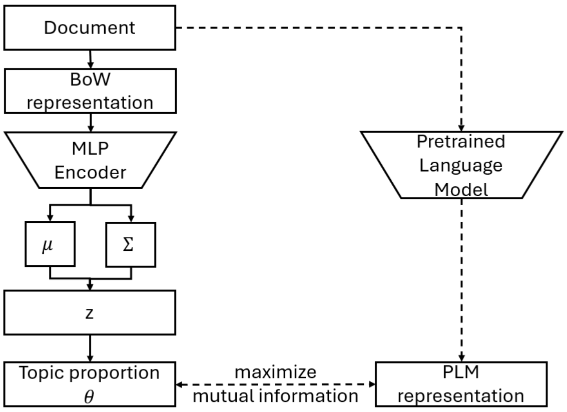

Document
BoW representation
MLP Encoder
μ
Σ
z
Topic proportion θ
maximize mutual information
Pretrained Language Model
PLM representation

Figure 1: High-level architecture of our encoder. Dashed line represent the part of our model that could be excluded in inference time.

2 Related Work

Topic Models and Neural Topic Models. Traditionally, generative probabilistic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) have been utilized for topic modeling. Numerous extensions of these models have been proposed to accommodate various assumptions and settings (Duc et al., 2017; Nguyen et al., 2019; Van Linh et al., 2022; Li et al., 2015; Nguyen et al., 2022b; Blei and Lafferty, 2006; Bianchi et al., 2021b). Recent advancements have integrated topic models with Variational Autoencoders (VAE) (Kingma and Welling, 2013) to improve scalability and efficiency in the inference process (Srivastava and Sutton, 2017; Dieng et al., 2020; Wu et al., 2023a; Cvejoski et al., 2023; Wang et al., 2022; Nguyen et al., 2024; Bianchi et al., 2021b; Wu et al., 2024b). Due to the difficulty of sampling the Dirichlet prior using the reparameterization trick, a Laplacian approximation is utilized in (Srivastava and Sutton, 2017). An alternative method involves using rejection sampling variational inference for the Dirichlet prior (Burkhardt and Kramer, 2019).

Within VAE architecture, two lines of research aim at improving the two components of the model, the encoder and the decoder. In both directions, incorporating external knowledge like word embeddings has become a prevalent practice to enhance topic quality (Bach et al., 2023; Dieng et al., 2020; Bianchi et al., 2021a; Grootendorst, 2022; Sia et al., 2020). Regarding the decoder (or the reconstruction phase of topic modeling), (Dieng et al., 2020) proposed using word embeddings, such as

Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), to gain a better understanding of vocabulary semantics, thus creating topics with more semantically related words. Various variants of word embedding are considered; for example, (Xu et al., 2023) utilized spherical embeddings to improve clusterability, while (Xu et al., 2022) employed word embeddings in hyperbolic space for topic taxonomy. In the embedding space, dependency between words and topics is modeled as a similarity function, such as dot product (Dieng et al., 2020; Nguyen et al., 2022b), cosine similarity (Zhao et al., 2020), or exponent of negative L_2 distance (Wu et al., 2023b). Additionally, (Wu et al., 2023b) applied a clustering regularization technique to ensure that each topic embedding acts as the center of a distinct cluster of word embeddings, thereby mitigating the issue of topic collapse.

In terms of encoders, pretrained language models (Devlin et al., 2019; Radford and Narasimhan, 2018) are another type of external knowledge frequently incorporated. (Bianchi et al., 2021a,b) used contextualized document embedding from SBERT (Reimers and Gurevych, 2019) as input, capturing valuable information in complement to bag-of-word representation. (Han et al., 2023) further employed SBERT to generate term weights that are integrated into the reconstruction loss to filter out irrelevant words. These approaches lead to increased inference time for neural topic models, which poses challenges for real-time applications. An improvement that does not rely on external knowledge involves utilizing optimal transport distance to model the disparity between documents and topics (Wang et al., 2022; Zhao et al., 2020).

Along with the development of pretrained language models, an alternate line of research in topic modeling that does not employ a VAE-like architecture directly group document's embedding to generate topics (Grootendorst, 2022; Sia et al., 2020; Zhang et al., 2022). Although this approach is simpler and yields coherent topics, it is not trivial to infer the topic proportions for a document.

Mutual Information Maximization. Mutual information maximization has been extensively utilized in machine learning to develop representations that encapsulate the intrinsic structure of data (van den Oord et al., 2019; Hjelm et al., 2019). For example, (Guo et al., 2022) employed this method to mitigate catastrophic forgetting in continual learning, while (Radford et al., 2021) leveraged mutual information to align text embeddings with image embeddings. In the field of topic modeling, mutual information maximization has been applied to align topics across different languages (Wu et al., 2023a) and to derive meaningful document representations (Nguyen and Luu, 2021).

3 Background

Notations. $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{D}$ is a collection of D documents. $\mathbf{x}_{i\text{BoW}}, \mathbf{x}_{iPLM}$ are the corresponding bag-of-words representation and pretrained language model embeddig of document $\mathbf{x}_i$. V is the number of unique terms in our vocabulary. K is the number of desired topics to find. $\beta = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^{V \times K}$ denotes the topic-word distribution matrix. $\mathbf{W} \in \mathbb{R}^{V \times L}, \mathbf{T} \in \mathbb{R}^{K \times L}$ correspond to the word embeddings and topic embeddings, respectively. θ_i is the topic proportion of document $\mathbf{x}_i$. $\mathbb{1}_N$ is a vector of length N where every element is 1. $[\![n]\!]$ is the set of first n integers $\{1, 2, \ldots, n\}$. Δ^n is the probability simplex in $\mathbb{R}^n$: $\Delta^n = \{\theta \in \mathbb{R}^n | \theta_i \geq 0; \sum_{i=1}^{n} \theta_i = 1\}$. The inner product between two matrices A and B of the same size is represented as $\langle A, B \rangle = \sum_{i,j} A_{ij} B_{ij}$. The inner product between two vectors is defined similarly. $H(P) = -\langle P, \log P - 1 \rangle = -\sum_{i,j} P_{ij}(\log P_{ij} - 1)$ is the Shannon entropy of P. The KL divergence between P and Q is defined as $\text{KL}(P \| Q) = \sum_{i,j} P_{ij} \log \left(\frac{P_{ij}}{Q_{ij}}\right) - 1$.

The objective of neural topic modeling is to identify K latent topics within $\mathbf{X}$. Each topic is represented as a multinomial probability distribution over the V vocabulary words, resulting in a topic-word distribution matrix $\beta \in \mathbb{R}^{V \times K}$. The matrix β is then decomposed into two components: word embeddings and topic embeddings (Dieng et al., 2020; Xu et al., 2022). (Wu et al., 2023b) defined the decomposition 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 word embeddings such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014).

Another objective of neural topic models is to infer topic proportions for a document $\mathbf{x}_i$. (Srivastava and Sutton, 2017; Bianchi et al., 2021a; Dieng et al., 2020; Wu et al., 2023b) employ a VAE-like architecture. Specifically, the topic proportion θ

depends on a latent variable $\mathbf{z}$, which conforms to a logistic-normal distribution characterized as $p(z) = \mathcal{LN}(z|\mu_0, \Sigma_0)$. When considering a document $\mathbf{x}_i$, its BoW representation, $\mathbf{x}_{i\text{BoW}}$, is subjected to encoding through neural networks. These networks furnish the parameters of a normal distribution, with mean $\mu_i = h_\mu(\mathbf{x}_{i\text{BoW}})$ and diagonal covariance matrix $\Sigma_i = \text{diag}(h_\Sigma(\mathbf{x}_{i\text{BoW}}))$. Leveraging the reparameterization trick (Kingma and Welling, 2013), z_i is sampled from the posterior distribution $q(z_i|\mathbf{x}_i) = \mathcal{N}(z_i|\mu_i, \Sigma_i)$ following $z_i = \mu_i + \Sigma_i\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$. Subsequently, softmax function is applied to z, yielding topic proportion $\theta_i = \text{softmax}(z_i)$. Following this, the Bag-of-Words representation is reconstructed with the topic-word distribution matrix β from a multinomial distribution $\hat{\mathbf{x}}_{i\text{BoW}} \sim \text{Multi}(\text{softmax}(\beta\theta_i))$. This comprehensive process is instrumental in achieving our objective function for topic modeling, which consists of a reconstruction term and a regularization term as follows:

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

Other preliminary on mutual information maximization and entropic regularized optimal transport can be found in Appendix A.3.

4 Proposed Method

We enhance both the inference network (encoder) and the generative model (decoder) of neural topic models through mutual information maximization and group topic regularization, respectively. The details will be presented in the subsequent subsections.

4.1 Maximize Mutual Information with Pretrained Language Model

We design an architecture that preserves the knowledge from a pretrained language model (PLM) in the encoder, even after the PLM is removed. Our approach is based on the assumption that the embeddings from the pretrained language model and the topic proportions should exhibit high mutual information. Specifically, let $\mathbf{X}_{\text{PLM}}$ denote the distribution of the embeddings from the pretrained language model and Θ represent the distribution of topic proportions of the documents. The desired property can be achieved by maximizing the mutual information between these two random variables, $I(\mathbf{X}_{\text{PLM}}; \Theta)$.

For tractability, we alternatively maximize its lower bound (van den Oord et al., 2019):

$$I(\mathbf{X}_{\text{PLM}}; \Theta) \geq \log B + \frac{1}{D}\sum_{i=1}^{D}\log\frac{e^{f(\theta_i, \mathbf{x}_{i\text{PLM}})}}{\sum_{\theta' \in B_i} e^{f(\theta', \mathbf{x}_{i\text{PLM}})}}$$

where B_i is a set containing sampled topic proportions of document i, including a positive example and negative examples for $\mathbf{x}_{i\text{PLM}}$. B_i is chosen to be the set of topic proportions of documents in the same batch as $\mathbf{x}_i$, and therefore has a size of B. The function $f(\theta, x_{\text{PLM}})$ quantifies the similarity between the topic proportion θ and the PLM's embedding x_{PLM}. Specifically, we use $f(a, b) = \frac{\langle \phi_\theta(a), b\rangle}{\|\phi_\theta(a)\| \cdot \|b\|}$, where ϕ_θ are learnable linear projections. As B is chosen to be a constant, we therefore minimize the following InfoNCE loss:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{-1}{D}\sum_{i=1}^{D}\log\frac{e^{f(\theta_i, \mathbf{x}_{i\text{PLM}})}}{\sum_{\theta' \in B_i} e^{f(\theta', \mathbf{x}_{i\text{PLM}})}}$$

4.2 Group Topic Regularization

We now introduce a new topic regularization based on optimal transport (OT) (Peyré and Cuturi, 2020) for the decoder. Specifically, we assume that each topic contains an equal amount of information and conduct a process where information is transferred between topics based on their relationships, ensuring the total amount remains unchanged. This process helps us learn the relationship between topics. To make it easier to relate to the mass redistribution problem in OT (Peyré and Cuturi, 2020), we use the metaphor of K topics as K piles of soil, each with an equal mass of $\frac{1}{K}$. After transportation, the mass of each pile of soil remains $\frac{1}{K}$. The transportation cost between two topics is calculated based on the distance between them in the embedding space. The matrix C represents the transportation costs for all pairs of topics. The optimal transport plan Q reveals the relationships between topics.

Formally, let $C \in \mathbb{R}^{K\times K}$ be the cost matrix in Euclidean space for topic embeddings $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_K\}$. The transport plan Q is the solution to the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \langle Q, C\rangle - \epsilon H(Q) \\ \text{subject to} \quad & Q \in \mathbb{R}^{K\times K}, \\ & Q\mathbb{1}_K = Q^\top\mathbb{1}_K = \frac{1}{K}\mathbb{1}_K, \\ & Q_{i,i} = 0 \; \forall i \in [\![K]\!]. \end{aligned} \tag{1}$$

The regularization term $\epsilon H(Q)$ encourages the matrix Q to become dense, thereby facilitating the sharing of information across multiple topics (Blondel et al., 2018). To focus on the interrelationships between different topics, the constraint $Q_{i,i} = 0$ is imposed. In practice, we ensure that $Q_{i,i}$ remains sufficiently small by setting $C_{i,i}$ to a large value. Subsequently, the Sinkhorn algorithm is employed to solve the optimization problem (Cuturi, 2013).

Furthermore, given that documents are assumed to be delivered in groups with similar semantic meanings, it is assumed that topics also exhibit a cluster structure. To enforce this structure, we introduce a regularization term that aligns matrix Q with matrix P, which encodes the shared information between grouped topics, as follows:

$$\mathcal{L}_{\text{GR}} = \text{KL}\left(P \| Q\right) \tag{2}$$

We propose a method to construct the matrix $P \in \mathbb{R}^{K \times K}$ to represent the shared information between grouped topics aligning on Q. Our goal is to categorize the K topic embeddings into clusters that reflect closely related semantic relationships. In the initial training phases, word embeddings display minimal deviation from their initialized states, thereby maintaining most of their semantic associations and effectively guiding the development of the topic embeddings. Leveraging this semantic information, we employ the KMeans clustering method (MacQueen et al., 1967) to partition the K topics into G clusters. Subsequently, we establish the matrix $\hat{P}$ in the following manner:

$$\hat{P}_{ij} = \begin{cases} 1 & \text{if topics } i, j \text{ are the same cluster} \\ u & \text{otherwise} \end{cases}$$

where the hyperparameter $0 < u < 1$ controls the ratio of shared information between topics within different groups compared to those within the same groups. We construct the final predefined matrix P by normalizing $\hat{P}$ so that the elements in each row or column sum to $\frac{1}{K}$. The normalization process involves iteratively normalizing row-wise and projecting onto the space of symmetric matrices.

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:

$$\mathcal{L}_{\text{ECR}} = \sum_{i=1}^{V} \sum_{j=1}^{K} \|\mathbf{w}_i - \mathbf{t}_j\|^2 \pi^*_{ij} \tag{3}$$

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

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

where $C_{\text{WT}} \in \mathbb{R}^{V \times K}$ is the distance matrix between word embeddings and topic embeddings. π^* is obtained using the Sinkhorn algorithm (Cuturi, 2013). In summary, in addition to the topic model objective, we use three loss functions: $\mathcal{L}_{\text{ECR}}$ to capture word-topic relations, $\mathcal{L}_{\text{GR}}$ to regularize topic-topic relations, and $\mathcal{L}_{\text{InfoNCE}}$ to enhance the encoder.

We can now finalize our training process as a two-stage approach. The first stage aims to produce the matrix P for the group regularizer with the following objective function:

$$\begin{aligned} \mathcal{L}_{\text{stage}_1} = &\mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}} \mathcal{L}_{\text{ECR}} \\ &+ \lambda_{\text{InfoNCE}} \mathcal{L}_{\text{InfoNCE}} \end{aligned} \tag{5}$$

In practice, the first training stage requires only a few epochs to achieve effective topic groups. After obtaining P as described in Section 4.2, we proceed to the second stage using the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{stage}_2} = &\mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}} \mathcal{L}_{\text{ECR}} \\ &+ \lambda_{\text{GR}} \dot{\mathcal{L}}_{\text{GR}} + \lambda_{\text{InfoNCE}} \mathcal{L}_{\text{InfoNCE}} \end{aligned} \tag{6}$$

where $\lambda_{\text{GR}}, \lambda_{\text{InfoNCE}}, \lambda_{\text{ECR}}$ are weight hyperparameters. The full algorithm are described in Appendix A.1.

5 Experiments

5.1 Settings

Datasets. We employ 20 News Groups (20NG) (Lang, 1995), a popular benchmark for topic modeling, AGNews (Zhang et al., 2015), a corpus contains news articles from more than 2000 sources, IMDB (Maas et al., 2011), a dataset of movie reviews, Yahoo Answers (Yahoo) (Zhang et al., 2015) a dataset contains questions and answers from the Yahoo! Answer website, and BBC (Greene and

	20NG			BBC		
	NPMI	NPMI − In	$\mathrm{C_p}$	NPMI	NPMI − In	$\mathrm{C_p}$
LDA ‡	0.0057	0.0801	0.0727	-0.0746	-0.0199	-0.0684
ProdLDA ‡	-0.0158	-0.0610	-0.0429	0.0001	-0.0105	0.0647
ETM ‡	0.0052	0.1219	0.0527	-0.0212	0.0441	0.0829
CTM ‡	-0.0161	0.1244	-0.1415	0.0436	0.0714	0.2543
ClusterTM ‡	0.0135	-0.2870	0.0160	0.0255	0.0656	0.0588
BertTopic ‡	0.0609	-0.0903	0.2318	-0.0007	0.0943	0.0747
UTopic ‡	**0.1069**	0.1130	**0.4850**	0.0938	0.1256	**0.5388**
NeuroMax	0.0929	**0.1810**	0.4543	**0.1288**	**0.2174**	0.5310

Table 1: Topic coherence measures, for models containing 10 topics. Bold values and underlined values represent the best and second-best results, respectively. ‡ Results resported in (Han et al., 2023).

	20NG			BBC		
	NPMI	NPMI − In	$\mathrm{C_p}$	NPMI	NPMI − In	$\mathrm{C_p}$
LDA ‡	-0.0056	0.0661	0.0719	-0.0718	-0.0205	-0.0709
ProdLDA ‡	-0.0227	-0.0083	-0.0634	0.0084	0.0110	0.0569
ETM ‡	0.0234	0.0927	0.1207	-0.0333	0.0251	0.0416
CTM ‡	-0.0086	0.1149	0.0156	0.0289	0.1109	0.3254
ClusterTM ‡	0.0154	-0.2863	0.0082	0.0339	0.0990	0.0908
BertTopic ‡	0.0322	-0.0563	0.1515	0.0456	0.0762	0.2556
UTopic ‡	**0.0653**	**0.1231**	**0.3709**	0.0708	0.1018	0.3925
NeuroMax	0.0469	0.0904	0.3104	**0.0742**	**0.1432**	**0.3938**

Table 2: Topic coherence measures, for models containing 20 topics. Bold values and underlined values represent the best and second-best results, respectively. ‡ Results resported in (Han et al., 2023).

Cunningham, 2006) - a corpus from BBC news website in 2004 and 2005.

Evaluation Metrics. We follow the evaluation methodology proposed in (Wu et al., 2023b) to assess both the quality of topics and the quality of document-topic distributions. Topic quality is evaluated using measurements of topic coherence and topic diversity. For topic coherence, we employ $\mathrm{C_V}$, NPMI, and $\mathrm{C_p}$, which are established metrics in topic modeling known for their high correlation with human judgment (Röder et al., 2015). These coherence measures are computed using a version of the Wikipedia corpus[1] as an external reference corpus. The NPMI measure is also computed using the training dataset as a reference dataset (denoted as NPMI − In). For topic diversity, we use the proportion of unique words among the topic words. Document-topic distribution quality is evaluated using NMI and Purity (Manning et al., 2008) on the document clustering task.

Baseline models. The first line of topic models not incorporating pre-trained language models includes: **LDA**, a probabilistic topic model introduced by (Blei et al., 2003), **ProdLDA** (Srivastava and Sutton, 2017), a variant of the LDA model that integrates Variational Autoencoders (VAEs), **ETM** (Dieng et al., 2020), a neural topic model that incorporates word embeddings, **NSTM** (Zhao et al., 2020), which utilizes the Sinkhorn distance to model the discrepancy between document-word distributions and document-topic distributions, **WeTe** (Wang et al., 2022), which alternately employs a conditional transport distance, and **ECRTM** (Wu et al., 2023b), which implements clustering regularization to improve topic coherence and distinctiveness. A series of topic models leveraging pre-trained language models includes two that employ a VAE-like architecture: **CTM** (Bianchi et al., 2021a), which utilizes contextualized embeddings as inputs to the neural topic model to capture richer semantic information, and **UTopic** (Han et al., 2023), which integrates tf-idf into the reconstruction loss to filter out unrelated words in a topic. Additionally, there are two baseline models that do not use a VAE-like architecture: **ClusterTM** (Sia et al., 2020), which clusters documents based on their contextual embeddings and

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

	20NG				Yahoo			
	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
LDA †	0.385	0.655	0.367	0.364	0.359	0.843	0.288	0.144
ETM †	0.375	0.704	0.347	0.319	0.354	0.719	0.405	0.192
NSTM †	0.395	0.427	0.354	0.356	0.39	0.658	0.395	0.241
WeTe †	0.383	0.949	0.268	0.304	0.367	0.878	0.389	0.252
ECRTM †	0.431	**0.964**	0.560	0.524	0.405	**0.985**	0.550	0.295
Utopic	**0.508**	0.860	0.530	0.454	**0.468**	0.788	0.473	0.244
NeuroMax	0.435	0.912	**0.623**	**0.570**	0.404	0.979	**0.588**	**0.331**
	IMDB				AGNews			
	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
LDA †	0.347	0.788	0.614	0.041	0.364	0.864	0.64	0.193
ETM †	0.346	0.557	0.66	0.038	0.364	0.819	0.679	0.224
NSTM †	0.334	0.175	0.658	0.040	0.411	0.8773	0.7719	0.324
WeTe †	0.368	0.931	0.587	0.031	0.383	0.945	0.641	0.268
ECRTM †	0.393	**0.974**	0.694	0.058	0.466	**0.961**	0.802	0.367
Utopic	**0.429**	0.554	0.550	0.005	**0.545**	0.838	0.768	0.303
NeuroMax	0.402	0.936	**0.709**	**0.061**	0.385	0.952	**0.804**	**0.410**

Table 3: Topic quality, quantified by mean C_V and mean TD, and document-topic quality, evaluated using mean NMI and mean Purity, for a model containing 50 topics. Bold values and underlined values represent the best and second-best results, respectively. †Results reported in (Wu et al., 2023b).

	20NG				Yahoo			
	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
LDA †	0.387	0.622	0.364	0.346	0.359	0.602	0.297	0.148
ETM †	0.369	0.573	0.394	0.339	0.353	0.624	0.428	0.208
NSTM †	0.391	0.473	0.383	0.363	0.387	0.659	0.405	0.242
WeTe †	0.352	0.742	0.338	0.348	0.353	0.544	0.444	0.269
ECRTM †	0.405	0.904	0.555	0.494	0.389	0.903	0.563	0.311
Utopic	**0.523**	0.750	0.545	0.452	**0.476**	0.612	0.549	0.305
NeuroMax	0.412	**0.913**	**0.602**	**0.516**	0.390	**0.922**	**0.583**	**0.329**
	IMDB				AG News			
	C_V	TD	Purity	NMI	C_V	TD	Purity	NMI
LDA †	0.342	0.691	0.600	0.037	0.349	0.696	0.654	0.194
ETM †	0.341	0.371	0.648	0.037	0.371	0.773	0.674	0.204
NSTM †	0.34	0.255	0.659	0.039	0.421	0.832	0.764	0.359
WeTe †	0.293	0.638	0.589	0.025	0.363	0.827	0.699	0.271
ECRTM †	0.373	**0.887**	0.694	0.049	0.416	**0.981**	0.812	**0.428**
Utopic	**0.534**	0.656	0.553	0.004	**0.548**	0.681	0.760	0.283
NeuroMax	0.381	0.870	**0.706**	**0.059**	0.406	0.957	**0.828**	0.389

Table 4: Topic quality, quantified by mean C_V and mean TD, and document-topic quality, evaluated using mean NMI and mean Purity, for a model containing 100 topics. Bold values and underlined values represent the best and second-best results, respectively. †Results reported in (Wu et al., 2023b).

utilizes term frequency (tf) to generate topic words, and **BertTopic** (Grootendorst, 2022), which employs a class-based variation of tf-idf to generate topic representations.

5.2 Topic Quality and Doc-Topic Distribution Quality

We first conducted experiments to assess the efficacy of our approach in comparison to baseline methods utilizing a pre-trained language model. Specifically, we employed two datasets: 20 News

	20NG				Yahoo			
	$\mathrm{C_V}$	TD	Purity	NMI	$\mathrm{C_V}$	TD	Purity	NMI
ECRTM	0.431	**0.964**	0.560	0.524	0.405	**0.985**	0.55	0.295
NeuroMax	0.435	0.912	**0.623**	**0.570**	0.404	0.979	**0.588**	**0.331**
w/o GR	**0.437**	0.940	0.613	0.554	**0.410**	0.957	0.577	0.324
w/o InfoNCE	**0.437**	0.924	0.595	0.547	0.404	0.937	0.564	0.317

Table 5: Ablation study on 20NG and Yahoo datasets.

	20NG	IMDB	AGNews	Yahoo
UTopic	44.46	98.69	39.48	39.42
NeuroMax	**0.15**	**0.24**	**0.14**	**0.15**

Table 6: Inference time of UTopic and Ours for different datasets with 50 topics. Experiments conducted on a NVIDIA RTX 3060 GPU.

Groups and BBC News. We adhere to the preprocessing procedures outlined by (Han et al., 2023). Tables 1 and 2 present the results of three topic coherence measures for 10 and 20 topics, respectively. Our approach demonstrates performance comparable to that of UTopic and consistently outperforms other methods. This outcome is expected, as the integration of the PLM's knowledge does not directly maximize the mutual information between topic proportions and contextualized representations. Instead, it operates via a lower bound approximation, which leads to improved topic coherence compared to methods that do not rely on PLMs, but does not achieve comparable embedding quality to methods that directly utilize PLMs, resulting in suboptimal overall performance.

We subsequently conducted experiments to evaluate the overall topic quality and document-topic distribution quality across four datasets: 20NG, Yahoo, IMDB, and AG News. The bag-of-words representation was obtained following the preprocessing steps described in (Wu et al., 2023b), and the contextualized embeddings were obtained after removing newline characters. Tables 3 and 4 report the topic quality and document-topic distribution quality for 50 and 100 topics, respectively. We provide the descriptive statistic in Appendix A.4. UTopic, as discussed in the previous experiment, demonstrates superior topic coherence performance but performs worse in terms of document-topic distribution quality. Compared to other baselines, our method achieves comparable topic quality and superior clustering performance owing to the integrated contextualized information and group regularization, which enhance the distinguishability of topic groups. Moreover, we illustrate the topics and their relationships in Appendix A.5.

5.3 Ablation Study

We conducted an ablation study on the 20NG dataset to analyze the impact of each component of our model on overall performance. Specifically, we iteratively remove the group regularizer and the InfoNCE loss, subsequently evaluating the model's performance. Table 5 presents the results obtained. In terms of document-topic distribution quality, measured by NMI and Purity, both components enhance the quality of the distribution. The model incorporating both components achieves the highest performance. Regarding topic quality, the model's performance remains competitive even with the removal of one component.

5.4 Inference time

We conducted experiments to measure the inference time of our model compared to UTopic, a model that utilizes contextualized embeddings as input. The results, presented in Table 6, demonstrate that our method achieves approximately 300 times faster inference while maintaining competitive performance, as discussed in Section 5.2. These results highlight that the method of maximizing mutual information can effectively address the issue of high inference costs associated with pre-trained language models with an acceptable performance tradeoff.

6 Conclusion

In conclusion, this paper introduces NeuroMax, a novel framework designed to tackle critical challenges in neural topic modeling. By maximizing the mutual information between the topic representation obtained from the common encoder in neural topic models and the representation derived from the PLM and leveraging optimal transport to capture topic relationships, NeuroMax offers a comprehensive solution for improving topic modeling efficiency and quality. Our experimental results demonstrate that NeuroMax significantly reduces inference time and obtains more coherent topics and topic groups, thus enhancing document representation for downstream task effectiveness. With its innovative approach and promising results, NeuroMax represents a valuable contribution to the field of neural topic modeling.

Limitations

Our proposed method has several limitations. First is the necessity to predefine the number of topics and groups as hyperparameters. This requirement is undesirable in real-world applications where the number of topics and topic groups are needed to be determined dynamically. A potential solution is to utilize the stick-breaking process, as demonstrated in (Chen et al., 2021; Ning et al., 2020), which can automatically determine the number of topics necessary. Another limitation is the challenge of applying our method to other scenarios, particularly dynamic topic models, online learning, and streaming learning. Adapting our approach to effectively capture the relationships between topics in temporal data remains an area for future research.

References

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Tran Xuan Bach, Nguyen Duc Anh, Ngo Van Linh, and Khoat Than. 2023. Dynamic transformation of prior knowledge into bayesian models for data streams. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3742–3750.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. 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*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

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

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. 2018. Smooth and sparse optimal transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889. PMLR.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021. Tree-structured topic modeling with nonparametric neural variational inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2343–2353, Online. Association for Computational Linguistics.

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

Kostadin Cvejoski, Ramsés J. Sánchez, and César Ojeda. 2023. Neural dynamic focused topic model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12719–12727.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages

4171–4186. Association for Computational Linguistics.

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

Anh Nguyen Duc, Ngo Van Linh, Anh Nguyen Kim, and Khoat Than. 2017. Keeping priors in streaming bayesian learning. In *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*, pages 247–258. Springer.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 377–384, New York, NY, USA. Association for Computing Machinery.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yiduo Guo, Bing Liu, and Dongyan Zhao. 2022. Online continual learning through mutual information maximization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8109–8126. PMLR.

Cuong Ha, Van-Dang Tran, Linh Ngo Van, and Khoat Than. 2019. Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout. *International Journal of Approximate Reasoning*, 112:85–104.

Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung, and Meeyoung Cha. 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*, pages 1802–1817, Dubrovnik, Croatia. Association for Computational Linguistics.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. *Preprint*, arXiv:1808.06670.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery.

Liran Juan, Yongtian Wang, Jingyi Jiang, Qi Yang, Guohua Wang, and Yadong Wang. 2020. Evaluating individual genome similarity with a topic model. *Bioinformatics*, 36(18):4757–4764.

Pooja Kherwa and Poonam Bansal. 2019. Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24).

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

Hoa M. Le, Son Ta Cong, Quyen Pham The, Ngo Van Linh, and Khoat Than. 2018. Collaborative topic model for poisson distributed ratings. *International Journal of Approximate Reasoning*, 95:62–76.

Ximing Li, Jihong Ouyang, You Lu, Xiaotang Zhou, and Tian Tian. 2015. Group topic model: organizing topics into groups. *Information Retrieval Journal*, 18(1):1–25.

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

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Duc Anh Nguyen, Kim Anh Nguyen, Canh Hao Nguyen, Khoat Than, et al. 2021. Boosting prior knowledge in streaming variational bayes. *Neurocomputing*, 424:143–159.

Ha Nguyen, Hoang Pham, Son Nguyen, Ngo Van Linh, and Khoat Than. 2022a. Adaptive infinite dropout for noisy and sparse data streams. *Machine Learning*, 111(8):3025–3060.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems*, volume 34, pages 11974–11986. Curran Associates, Inc.

Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Topic modeling as multi-objective contrastive optimization. *Preprint*, arXiv:2402.07577.

Tung Nguyen, Trung Mai, Nam Nguyen, Linh Ngo Van, and Khoat Than. 2022b. Balancing stability and plasticity when learning topic models from short and noisy text streams. *Neurocomputing*, 505:30–43.

Tung Nguyen, Tung Pham, Linh Van Ngo, Ha-Bang Ban, and Khoat Quang Than. Out-of-vocabulary handling and topic quality control strategies in streaming topic models. *Available at SSRN 4592178*.

Van-Son Nguyen, Duc-Tung Nguyen, Linh Ngo Van, and Khoat Than. 2019. Infinite dropout for training bayesian models from data streams. In *2019 IEEE international conference on big data (Big Data)*, pages 125–134. IEEE.

Xuefei Ning, Yin Zheng, Zhuxi Jiang, Yu Wang, Huazhong Yang, Junzhou Huang, and Peilin Zhao. 2020. Nonparametric topic modeling with neural inference. *Neurocomputing*, 399:296–306.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Gabriel Peyré and Marco Cuturi. 2020. Computational optimal transport. *Preprint*, arXiv:1803.00567.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 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)*, pages 1728–1736, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *Preprint*, arXiv:1703.01488.

Anh Phan Tuan, Bach Tran, Thien Huu Nguyen, Linh Ngo Van, and Khoat Than. 2020. Bag of biterms modeling for short texts. *Knowledge and Information Systems*, 62(10):4055–4090.

Francisco B. Valero, Marion Baranes, and Elena V. Epure. 2022. Topic modeling on podcast short-text metadata. In *Advances in Information Retrieval*, pages 472–486, Cham. Springer International Publishing.

Hugues Van Assel, Titouan Vayer, Rémi Flamary, and Nicolas Courty. 2023. SNEkhorn: Dimension Reduction with Symmetric Entropic Affinities. In *Advances in Neural Information Processing Systems*, volume 36, pages 44470–44487. Curran Associates, Inc.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ngo Van Linh, Nguyen Kim Anh, Khoat Than, and Chien Nguyen Dang. 2017. An effective and interpretable method for document classification. *Knowledge and Information Systems*, 50(3):763–793.

Ngo Van Linh, Tran Xuan Bach, and Khoat Than. 2022. A graph convolutional topic model for short and noisy text streams. *Neurocomputing*, 468:345–359.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. *Preprint*, arXiv:2203.01570.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023a. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13763–13771.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization. *Preprint*, arXiv:2306.04217.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2023c. Towards the topmost: A topic modeling system toolkit. *arXiv preprint arXiv:2309.06908*.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024b. On the affinity, rationality, and diversity of hierarchical topic modeling. *Preprint*, arXiv:2401.14113.

Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457, Toronto, Canada. Association for Computational Linguistics.

Yi.shi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, and Mingyuan Zhou. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In *Advances in Neural Information Processing Systems*, volume 35, pages 31557–31570. Curran Associates, Inc.

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

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 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*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2020. Neural topic model via optimal transport. *Preprint*, arXiv:2008.13537.

A Appendix

A.1 Algorithm

The detail training algorithm for NeuroMax is presented in Algorithm 1.

A.2 Implementation Details.

Our implementation builds upon PyTorch (Ansel et al., 2024) and TopMost (Wu et al., 2023c, 2024a), a publicly available toolkit for topic modeling. Palmetto (Röder et al., 2015) is used to quantify topic coherence.We employ the $\mathrm{allMiniLM-L6-v2}$ model (Reimers and Gurevych, 2019) as our pre-trained language model. GloVe (Pennington et al., 2014) serves as the initial word embedding. Following the architecture in (Wu et al., 2023b), we utilize the same encoder network, comprising a two-layer softplus-activated MLP and an additional layer for the mean and covariance of the latent variable. We also train our model for $N = 500$ epochs with a batch size of 200, utilizing the Adam optimizer (Kingma and Ba, 2017) with a learning rate set to 0.002. The hyperparameter u of the group regularizer is set to $\frac{1}{5}$, and the number of first-stage training epochs is set to 10. The weight hyperparameters are searched in ranges as follows:

Parameter	Values
λ_{ECR}	$20, 40, 50, 60, 80, 100, 150, 200, 250$
λ_{GR}	$1, 5, 10, 20, 50$
λ_{InfoNCE}	$1, 10, 30, 50, 80, 100, 130, 150$

Table 7: Value range for hyperparemeter searching

A.3 Preliminary

A.3.1 Mutual Infomation Maximization

Let X and Y be two random variables. The mutual information between X and Y, which quantifies the degree of dependence between the two variables, is defined as:

$$I(X;Y) = \int_X \int_Y p(X,Y) \log \frac{p(X,Y)}{p(X)p(Y)} dxdy \tag{7}$$

In general, directly maximizing this quantity is intractable. We resort to its lower bound for tractable maximization (van den Oord et al., 2019):

$$\begin{aligned} I(X;Y) &\geq \mathcal{L}_{\text{InfoNCE}} \\ &= \log N + \mathbb{E}_{p(x,y)} \left[\log \frac{f(x,y)}{\sum_{y' \in B} f(x,y')} \right] \end{aligned} \tag{8}$$

where $f(x,y)$ is a similarity score between x and y, and B is the set containing one positive and $N-1$ negative examples. Intuitively, maximizing this lower bound encourages a data instance to have a high similarity score with its positive example and a low similarity score with its negative examples.

A.3.2 Entropic Regularized Optimal Transport

Let u and v be two discrete measures on the supports $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$ and $\{y_1, y_2, \dots, y_m\} \subset \mathbb{R}^d$, respectively, with associated weights $(u_1, u_2, \dots, u_n)$ and $(v_1, v_2, \dots, v_m)$ satisfying $\sum_i u_i = \sum_j v_j$. Given a cost matrix $C \in \mathbb{R}^{n \times m}$, an optimal transport plan is defined as the solution to the following optimization problem (Peyré and Cuturi, 2020):

Algorithm 1 Learning NeuroMax

```
Input: Document collection $\mathbf{X}$, pretrained language model PLM, pretrained word embedding $\mathbf{W}_{\text{pretrained}}$, number of topic $K$,
  total number of training epoch $N$, number of training epochs for the first stage $M$;
Output: Encoder network's parameter $W_{\text{enc}}$, linear projections' parameter $W_{\phi_\theta}$, word embedding $\mathbf{W}$, topic embedding $\mathbf{T}$,
  word-topic transport plan $\pi^*$, topic-topic transport plan $Q$;
  Initialize $\mathbf{W} = \mathbf{W}_{pretrained}$
  for $t = 1, 2, \ldots, N$ do
    // Update parameters related to the encoder
    Update $W_{\text{enc}}$ and $W_{\phi_\theta}$ using a gradient descent step based on the loss $\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_{\text{InfoNCE}}\mathcal{L}_{\text{InfoNCE}}$
    // Update parameters related to the decoder
    if $t \leq M$ then
      // Stage 1
      Calculate word-topic distance matrix $C_{WT}$
      Update $\pi^*$ as the solution of problem (4) by Sinkhorn algorithm
      Update $\mathbf{T}$, $\mathbf{W}$ using a gradient descent step based on the loss $\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}}\mathcal{L}_{\text{ECR}}$
      if $t = M$ then
        Calculate matrix $P$
      end if
    else
      // Stage 2
      Calculate word-topic distance matrix $C_{WT}$ and topic-topic distance matrix $C$
      Update $Q$ and $\pi^*$ using Sinkhorn algorithm to sovle the OT problems (1) and (4) respectively
      Update $\mathbf{T}$, $\mathbf{W}$ using a gradient descent step based on $\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_{\text{ECR}}\mathcal{L}_{\text{ECR}} + \lambda_{\text{GR}}\mathcal{L}_{\text{GR}}$
    end if
  end for
```

$$\begin{aligned}&\underset{P\in\mathbb{R}^{n\times m}}{\text{minimize}}\ \langle P, C\rangle\\ &\text{subject to } P\mathbb{1} = u, P^\top\mathbb{1} = v.\end{aligned}\tag{9}$$

The minimized objective function of the aforementioned problem can serve as a measure of the distance between two distributions. This distance, along with the optimal plan, can be approximated efficiently by solving a modified problem, specifically by introducing an entropic regularization term to (9) and employing iterative algorithms such as Sinkhorn algorithm (Cuturi, 2013).

The entropic regularization encourages the optimal transport plan to be dense (Blondel et al., 2018). Consequently, if $u \equiv v$, mass from a given atom is compelled to disperse to other atoms, with a higher proportion of mass allocated to nearer atoms. Therefore, the entropic regularized optimal transport plan can be utilized as a similarity measure for data points (Van Assel et al., 2023).

A.4 Descriptive Statistic

In Tables 8 and 9, we report the performance of the NeuroMax model, providing both the mean and standard deviation over five independent runs. For comparison, the results for the UTopic model are also included as a reference baseline.

Table 8: Comparison of UTopic and NeuroMax on 50 topics (mean ± std)

Dataset	Metric	UTopic	NeuroMax
20NG	C_V	0.508 ± 0.006	0.435 ± 0.004
	TD	0.860 ± 0.032	0.912 ± 0.045
	Purity	0.530 ± 0.010	0.623 ± 0.022
	NMI	0.454 ± 0.003	0.570 ± 0.014
IMDB	C_V	0.429 ± 0.014	0.402 ± 0.005
	TD	0.554 ± 0.045	0.936 ± 0.025
	Purity	0.550 ± 0.004	0.709 ± 0.001
	NMI	0.005 ± 0.001	0.061 ± 0.002
Yahoo	C_V	0.468 ± 0.012	0.404 ± 0.002
	TD	0.788 ± 0.007	0.979 ± 0.003
	Purity	0.473 ± 0.009	0.588 ± 0.004
	NMI	0.244 ± 0.008	0.331 ± 0.002
AGNews	C_V	0.545 ± 0.008	0.385 ± 0.007
	TD	0.838 ± 0.025	0.952 ± 0.026
	Purity	0.768 ± 0.018	0.804 ± 0.006
	NMI	0.303 ± 0.012	0.410 ± 0.007

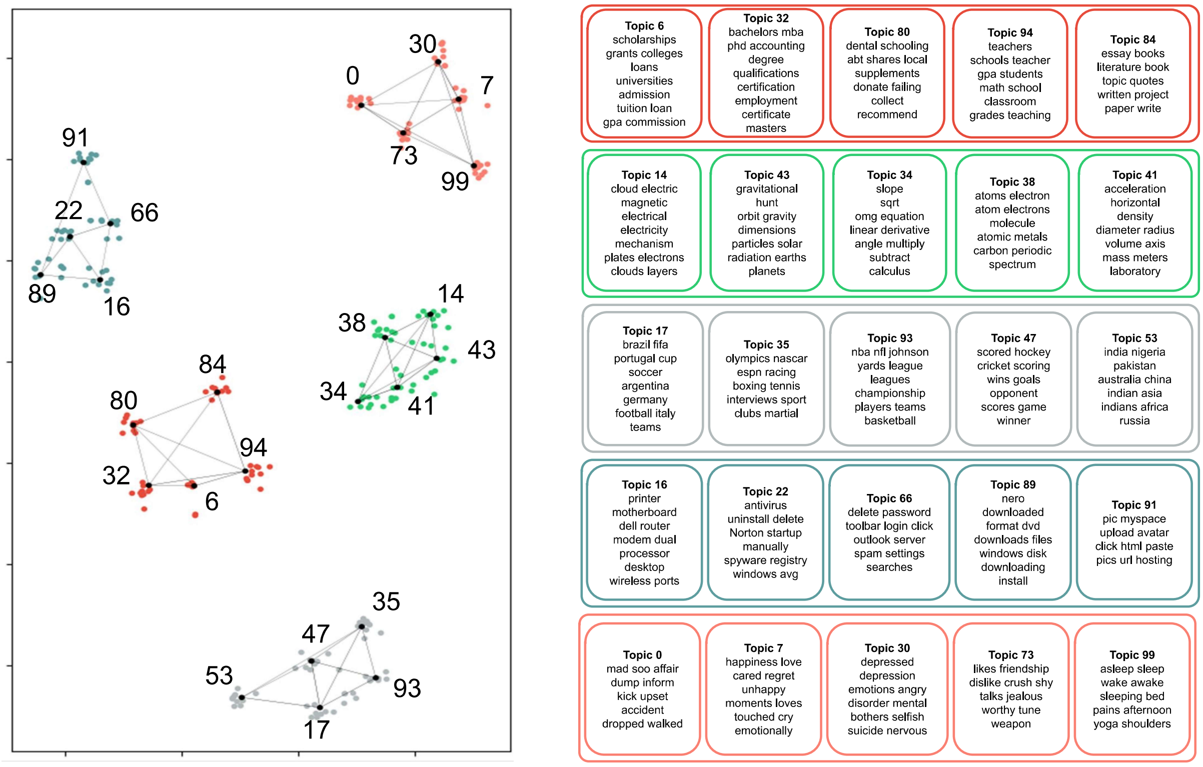
A.5 Topic visualization

To assess the effectiveness of our grouping regularization, we visualized word and topic embeddings of 5 randomly selected groups, each comprising 5 randomly chosen topics, and displayed the corresponding topic words in Figure 2. We observed that topics within the same group tend to share more information (highlighted in gray) and share semantically similar words, while topics from different groups display distinct words and lower sharing scores. This highlights the efficacy of our group regularizer in generating closely embedded, semantically similar topics. Furthermore,

Table 9: Comparison of UTopic and NeuroMax on 100 topics (mean ± std)

Dataset	Metric	UTopic	NeuroMax
20NG	$\mathrm{C_V}$	0.523 ± 0.006	0.412 ± 0.003
	TD	0.750 ± 0.012	0.913 ± 0.002
	Purity	0.545 ± 0.006	0.602 ± 0.007
	NMI	0.452 ± 0.006	0.516 ± 0.005
IMDB	$\mathrm{C_V}$	0.534 ± 0.004	0.381 ± 0.005
	TD	0.656 ± 0.028	0.870 ± 0.027
	Purity	0.553 ± 0.003	0.706 ± 0.004
	NMI	0.004 ± 0.001	0.059 ± 0.003
Yahoo	$\mathrm{C_V}$	0.476 ± 0.013	0.390 ± 0.002
	TD	0.612 ± 0.044	0.922 ± 0.029
	Purity	0.549 ± 0.014	0.583 ± 0.005
	NMI	0.305 ± 0.010	0.329 ± 0.003
AGNews	$\mathrm{C_V}$	0.548 ± 0.003	0.406 ± 0.007
	TD	0.681 ± 0.021	0.957 ± 0.015
	Purity	0.760 ± 0.011	0.828 ± 0.010
	NMI	0.283 ± 0.011	0.389 ± 0.014

the topics within the same group are not collapsing, thanks to the ECR regularizer.



Topic 6 scholarships grants colleges loans universities admission tuition loan gpa commission	**Topic 32** bachelors mba phd accounting degree qualifications certification employment certificate masters	**Topic 80** dental schooling abt shares local supplements donate failing collect recommend	**Topic 94** teachers schools teacher gpa students math school classroom grades teaching	**Topic 84** essay books literature book topic quotes written project paper write
Topic 14 cloud electric magnetic electrical electricity mechanism plates electrons clouds layers	**Topic 43** gravitational hunt orbit gravity dimensions particles solar radiation earths planets	**Topic 34** slope sqrt omg equation linear derivative angle multiply subtract calculus	**Topic 38** atoms electron atom electrons molecule atomic metals carbon periodic spectrum	**Topic 41** acceleration horizontal density diameter radius volume axis mass meters laboratory
Topic 17 brazil fifa portugal cup soccer argentina germany football italy teams	**Topic 35** olympics nascar espn racing boxing tennis interviews sport clubs martial	**Topic 93** nba nfl johnson yards league leagues championship players teams basketball	**Topic 47** scored hockey cricket scoring wins goals opponent scores game winner	**Topic 53** india nigeria pakistan australia china indian asia indians africa russia
Topic 16 printer motherboard dell router modem dual processor desktop wireless ports	**Topic 22** antivirus uninstall delete Norton startup manually spyware registry windows avg	**Topic 66** delete password toolbar login click outlook server spam settings searches	**Topic 89** nero downloaded format dvd downloads files windows disk downloading install	**Topic 91** pic myspace upload avatar click html paste pics url hosting
Topic 0 mad soo affair dump inform kick upset accident dropped walked	**Topic 7** happiness love cared regret unhappy moments loves touched cry emotionally	**Topic 30** depressed depression emotions angry disorder mental bothers selfish suicide nervous	**Topic 73** likes friendship dislike crush shy talks jealous worthy tune weapon	**Topic 99** asleep sleep wake awake sleeping bed pains afternoon yoga shoulders

Figure 2: (Left) t-SNE visualization of topics embeddings (black dots) and embeddings of their top 10 word (color dots). Word embeddings for topics within the same group share the same color. Pairs of topics with high information sharing scores are highlighted in gray. (Right) Corresponding top 10 words for each topic.