HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Embedded topic models are able to learn interpretable topics even with large and heavy-tailed vocabularies. However, they generally hold the Euclidean embedding space assumption, leading to a basic limitation in capturing hierarchical relations. To this end, we present a novel framework that introduces hyperbolic embeddings to represent words and topics. With the tree-likeness property of hyperbolic space, the underlying semantic hierarchy among words and topics can be better exploited to mine more interpretable topics. Furthermore, due to the superiority of hyperbolic geometry in representing hierarchical data, the tree-structure knowledge can be naturally injected to guide the learning of a topic hierarchy. Therefore, we further develop a regularization term based on the contrastive learning concept to efficiently inject prior structural knowledge. Experiments on both topic taxonomy discovery and document representation tasks demonstrate the proposed framework achieves improved performance on the basis of existing embedded topic models.

1 Introduction

With a long track record of success in a variety of applications [1–6], topic models have been one of the most powerful tools for automatic text analysis. Typically, given a collection of documents, a topic model aims to identify a group of salient topics by capturing common word co-occurrence patterns. Despite their popularity, traditional topic models such as Latent Dirichlet Allocation (LDA) [7] and its variants [8–12] are plagued by the complicated posterior inference, presenting a challenge to create deeper and more expressive models of text. Fortunately, recent developments of Variational AutoEncoders (VAEs) and Autoencoding Variational Inference (AVI) [13, 14] have shed light on this problem, resulting in the proposal of a series of Neural Topic Models (NTMs) [15–18]. With better flexibility and scalability, NTMs have gained increasing research interest over the past few years.

Parallel to neural topic modeling, the idea of bringing word embeddings [19, 20] into topic models has also attracted much attention. Considering the large performance degradation over short texts due to limited word co-occurrence information, some early works [21–23] exploit word embeddings as complementary metadata and incorporate them into the generative process of topic models. Recently, more flexible ways of combining word embeddings have been explored thanks to the development of NTMs. For example, Bianchi et al. [24] use word embeddings directly as part of the encoder's input. In particular, a novel one called Embedded Topic Model (ETM) [25] stands out for its impressive performance as well as the elegant way it integrates word embeddings. Specifically, by representing topics as points in the word embedding space, ETM assigns probabilities to words based on their (inner product) distances from each topic embedding. As a result, semantically related words tend to fall near the same topic, thus facilitating the discovery of more interpretable topics.

Under the inspiration of ETM, Duan et al. [26] have extended a similar idea to hierarchical topic modeling and proposed the SawETM. In addition to mapping words and hierarchical topics into a

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

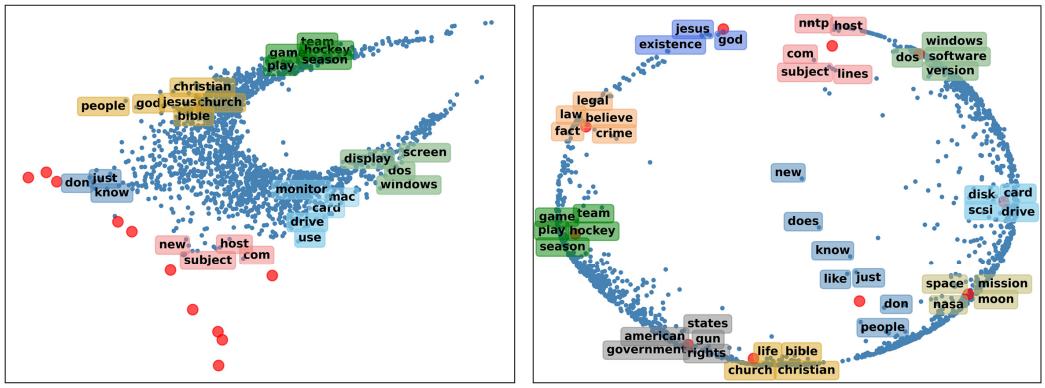

Figure 1: Visualization of 2D Euclidean embedding space (left) and 2D hyperbolic embedding space (right) learned by ETM. Red points denote topic embeddings, blue points represent word embeddings.

shared embedding space, SawETM has also developed a unique Sawtooth Connection module to capture the dependencies between the topics at different layers, which, on the other side, empowers it to support a deep network structure. While achieving promising results, both ETM and SawETM hold the Euclidean embedding space assumption, leading to a fundamental limitation that their ability to model complex patterns (akin to social networks, knowledge graphs, and taxonomies) is inherently bounded by the dimensionality of the embedding space [27, 28]. As a consequence, the underlying semantic hierarchy among the words and topics can hardly be expressed adequately in a relatively low-dimensional embedding space, as illustrated in the left part of Figure 1.

Apart from the difficulty in capturing the implicit semantic hierarchy, another concomitant problem is the dilemma of incorporating explicit structural knowledge. Assuming we have a prior taxonomy of concepts and wish to use it to guide the learning of hierarchical topics, it is challenging to preserve the structure between concepts in Euclidean space by constraining the word and topic embeddings. To cope with this issue, TopicNet [29] employs the Gaussian-distributed embeddings as a substitute for the vector embeddings to represent words and topics. As such, the prior knowledge of hypernym relations between concepts could be naturally injected via the encapsulation of probability densities. However, maintaining the semantic hierarchy in such an embedding space still suffers from a certain degree of distortion, as it poses a tough to the optimization of KL divergence between distributions. Furthermore, the introduction of Gaussian-distributed embeddings entails a great demand on memory, limiting its potential scalability to large vocabularies and high-dimensional embedding spaces.

To overcome the above shortcomings brought by Euclidean embedding space, we propose to compute embeddings in hyperbolic space. Distinguished by the tree-likeness properties [30, 31], hyperbolic space has been consistently shown to be superior in modeling hierarchical data compared to Euclidean space [32–35]. Therefore, by measuring the distance between words and topics in the hyperbolic embedding space, the model is encouraged to better capture the underlying semantic hierarchy. As depicted in the right part of Figure 1, some universal words such as "new" and "does" fall around the center, they stay close to all other points because they often appear together with other words. While more specific words fall onto the boundary and are only close to the nearby words. Moreover, the hyperbolic space also provides a better platform to inject the prior structural knowledge, as the hierarchical relations can be effectively preserved by imposing constraints on the distance between word and topic embeddings. In a nutshell, the main contributions of this paper are as follows:

- We introduce the hyperbolic embedding space to improve existing embedded topic models, which is beneficial to both the mining of implicit semantic hierarchy and the incorporation of explicit structural knowledge.
- We design a novel node-level representation learning scheme based on the contrastive idea to effectively inject the prior structural knowledge to guide the learning of a meaningful topic taxonomy.
- Extensive experiments on both topic quality and document representation demonstrate that the proposed method achieves competitive performance against baselines.

2 Background

2.1 Embedded topic model

ETM [25] is a neural topic model that builds on two main techniques: LDA [7] and word embeddings [19, 20]. To marry the probabilistic topic modeling of LDA with the contextual information brought by word embeddings, ETM maintains vector representations of both words and topics and uses them to derive the per-topic distribution over the vocabulary. Specifically, consider a corpus with V distinct terms comprising the vocabulary, we denote the word embedding matrix as $\boldsymbol{\rho} \in \mathbb{R}^{D \times V}$, where D is the dimensionality of the embedding space. For each topic, there is also an embedding representation $\boldsymbol{\alpha}_k \in \mathbb{R}^D$, then ETM defines the per-topic distribution $\boldsymbol{\beta}_k$ over the vocabulary as

$$\boldsymbol{\beta}_k = \text{Softmax}\left(\boldsymbol{\rho}^\top \boldsymbol{\alpha}_k\right) \tag{1}$$

With the above definition, ETM specifies a generative process analogous to LDA. Let $\boldsymbol{w}_{jn} \in \{1, ..., V\}$ denote the nth word in the jth document, the generative process is as follows.

1. Draw topic proportions $\boldsymbol{\theta}_j \sim \mathcal{LN}\left(\mathbf{0}, \boldsymbol{I}\right)$.
2. For each word n in the document:
   (a) Draw topic assignment $\boldsymbol{z}_{jn} \sim \text{Cat}\left(\boldsymbol{\theta}_j\right)$.
   (b) Draw word $\boldsymbol{w}_{jn} \sim \text{Cat}\left(\boldsymbol{\beta}_{\boldsymbol{z}_{jn}}\right)$.

Where $\mathcal{LN}(\cdot)$ in step 1 denotes the logistic-normal distribution [36], which transforms a standard Gaussian random variable to the simplex. By taking the inner product of the word embedding matrix $\boldsymbol{\rho}$ and the topic embedding $\boldsymbol{\alpha}_k$ to derive $\boldsymbol{\beta}_k$, the intuition behind ETM is that semantically related words will be assigned to similar topics. With this property, ETM has been demonstrated to improve the quality of the learned topics, especially in the presence of large vocabularies. Like most NTMs, ETM is fitted via an efficient amortized variational inference algorithm.

2.2 Hyperbolic geometry

In this part, we briefly review some key concepts on hyperbolic geometry. A comprehensive and in-depth description can be found in Lee [37] and Nickel and Kiela [38]. Under the mathematical framework of Riemannian geometry, hyperbolic geometry specializes in the case of constant negative curvatures. Intuitively, the hyperbolic space can be understood as a continuous version of trees: the volume of a ball expands exponentially with its radius, just as how the number of nodes in a binary tree grows exponentially with its depth. Mathematically, there exist multiple equivalent models for hyperbolic space with different definitions and metrics. Here, we consider two representative ones in light of optimization simplicity and stability: Poincaré ball model [27] and the Lorentz model [38].

Poincaré ball model The Poincaré ball model of an n-dimensional hyperbolic space with curvature C ($C < 0$) is defined by the Riemannian manifold $\mathcal{P}^n = (\mathcal{B}^n, g_p)$, where $\mathcal{B}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < 1/\sqrt{-C}\}$ is the open n-dimensional ball with radius $1/\sqrt{-C}$ and g_p is the metric tensor that can be converted from the Euclidean metric tensor $g_e = I$.

$$g_p\left(\boldsymbol{x}\right) = \left(\frac{2}{1 + C\|\boldsymbol{x}\|^2}\right)^2 g_e \tag{2}$$

Lorentz model The Lorentz model (also named Hyperboloid model) of an n-dimensional hyperbolic space with curvature C ($C < 0$) is defined by the Riemannian manifold $\mathcal{L}^n = (\mathcal{H}^n, g_l)$, where $\mathcal{H}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = 1/C\}$ and $g_l = \text{diag}\left([-1, \mathbf{1}_n^\top]\right)$. $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denote the *Lorentzian inner product*. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n+1}$, the Lorentz inner product induced by g_l is calculated as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = \boldsymbol{x}^\top g_l \boldsymbol{y} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i \tag{3}$$

Some other related formulas and operations in hyperbolic space, e.g., distance, maps, and parallel transport will be further introduced in Appendix. C.

3 Hyperbolic Enhanced Hierarchical Topic Modeling

In this section, we elaborate on how the introduced hyperbolic embeddings benefit the mining of implicit semantic hierarchy and the incorporation of explicit tree-structure knowledge, both of which encourage the model to find more interpretable topics.

3.1 Topic modeling in hyperbolic space

Theoretically, the idea of representing words and topics in hyperbolic space is orthogonal to a wide range of topic models employing the word embeddings technique. To provide the foundation for the subsequent injection of structural knowledge, we here apply our method to a hierarchical embedded topic model SawETM [26]. SawETM utilizes the adapted Poisson gamma belief network (PGBN) [39] as its generative module (decoder) and decomposes the topic matrices into the inner product of topic embeddings at adjacent layers. The novelty of our method lies in that the hierarchical relations can be better reflected by the distances between embeddings in hyperbolic space. Mathematically, the generative model with L latent layers is formulated as

$$\begin{aligned}
&\boldsymbol{\theta}_j^{(L)} \sim \text{Gam}\left(\boldsymbol{\gamma}, e_j^{(L+1)}\right), \boldsymbol{\theta}_j^{(l)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)}, e_j^{(l+1)}\right), \ldots, \\
&\boldsymbol{\theta}_j^{(1)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, e_j^{(2)}\right), \boldsymbol{x}_j \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \\
&\boldsymbol{\Phi}^{(l)} = \text{Softmax}\left(\mathcal{S}\left(\boldsymbol{\alpha}^{(l-1)}, \boldsymbol{\alpha}^{(l)}\right)\right),
\end{aligned} \tag{4}$$

The above formula clearly describes how the multi-layer document representation is generated via a top-down process. Specifically, the latent representation of the top layer $\boldsymbol{\theta}_j^{(L)}$ is sampled from a fixed gamma prior distribution, then at each intermediate layer l the latent units $\boldsymbol{\theta}_j^{(l)} \in \mathbb{R}^{K_l}$ are factorized into the product of the factor loading matrix $\boldsymbol{\Phi}^{(l+1)} \in \mathbb{R}^{K_l \times K_{l+1}}$ and latent units $\boldsymbol{\theta}_j^{(l+1)} \in \mathbb{R}^{K_{l+1}}$ of the above layer. Until the bottom layer, the observation of word count vector $\boldsymbol{x}_j \in \mathbb{Z}^V$ is modeled as the Poisson distribution. Note that the subscript j denotes the document index and some other variables $\boldsymbol{\gamma}, e_j^{(L+1)}, \ldots, e_j^{(2)}$ are hyperparameters. Especially, the factor loading matrix $\boldsymbol{\Phi}^{(l)}$ of layer l is derived based on the distance between the topic embeddings at two adjacent layers, i.e., $\boldsymbol{\alpha}^{(l-1)} \in \mathbb{R}^{K_{l-1} \times D}$ and $\boldsymbol{\alpha}^{(l)} \in \mathbb{R}^{K_l \times D}$. Note that $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{V \times D}$ represents the word embeddings. Sine all embeddings are projected into the hyperbolic space to fully explore the underlying semantic hierarchy among the words and topics, we design our similarity score function as

$$\begin{aligned}
\mathcal{S}(\boldsymbol{x}, \boldsymbol{y}) &= -d_{\mathcal{P}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{-1}{\sqrt{|C|}} \text{arcosh}\left(1 - \frac{2C\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\left(1 + C\|\boldsymbol{x}\|^2\right)\left(1 + C\|\boldsymbol{y}\|^2\right)}\right) \\
\mathcal{S}(\boldsymbol{x}, \boldsymbol{y}) &= -d_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{-1}{\sqrt{|C|}} \text{arcosh}\left(\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}\right)
\end{aligned} \tag{5}$$

Where $d_{\mathcal{P}}(\cdot, \cdot)$ and $d_{\mathcal{L}}(\cdot, \cdot)$ are the distance functions of the Poincaré ball model and the Lorentz model, respectively. As the two models of hyperbolic space are mathematically equivalent, we take the Poincaré ball as an example for analysis. Eq. (5) shows that the distance changes smoothly with respect to the norm of $\boldsymbol{x}$ and $\boldsymbol{y}$. This locality plays a crucial role in learning continuous embeddings of hierarchies. For instance, the origin of $\mathcal{B}^n$ has a zero norm, it would have relatively small distance to all other nodes, which exactly corresponds to the root node of a tree. On the other hand, those points close to the boundary of the ball have a norm close to one, so the distance between them grows quickly, which properly reflects the relationships between the leaf nodes of a tree.

3.2 Knowledge injection with hyperbolic contrastive loss

Hierarchical structures are ubiquitous in knowledge representation and reasoning [38]. Particularly, mining a set of meaningful topics organized into a hierarchy from massive text corpora is intuitively appealing, as it allows users to easily access the information of their interest. However, most existing hierarchical topic models struggle to realize this goal without any supervision, and some appropriate guidance with structural prior knowledge proves to be helpful for mitigating this issue [29].

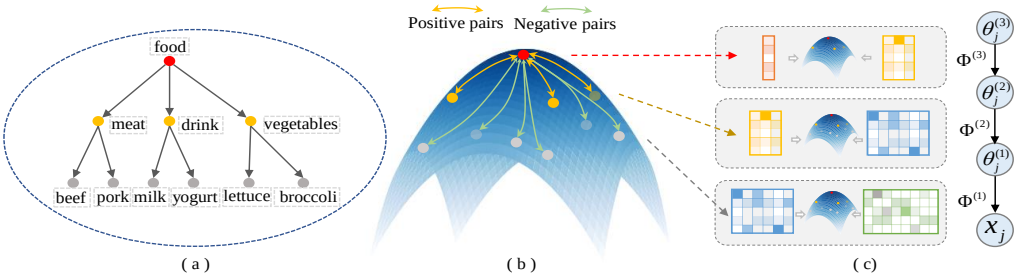

Figure 2: Overview of knowledge guided topic taxonomy discovery. (a) The prior concept taxonomy constructed from vocabulary; (b) Illustration of the strategy for picking positive pairs and negative pairs in the hyperbolic embedding space; (c) The hierarchical generative model whose factor loading matrices are derived based on the hyperbolic distances between topic embeddings at two adjacent layers.

Form of prior knowledge We assume the prior knowledge takes the form of a concept taxonomy, which is compatible with the deep structure of the proposed hierarchical topic model in Section 3.1. In detail, the taxonomy exhibits a top-down reification process and concepts between two adjacent layers with connections follow the hypernym relation, as shown in Figure 2(a). Meanwhile, to keep the taxonomy consistent with the corresponding dataset, we construct it by traversing each word in the vocabulary to find its ancestors along the hypernym paths provided by WordNet [40].

Hyperbolic contrastive loss Although the paradigm of contrastive learning has been successfully applied to graph representation learning in Euclidean space [41–44], those developed contrastive algorithms are not directly applicable in our case. On the one hand, we focus more on learning node representations while not destroying their prior hierarchical structures. On the other hand, hyperbolic space possesses distinctive properties (e.g, hierarchical awareness and spacious room) compared to its Euclidean counterpart. Consequently, to accommodate the two differences, we design a node-level hyperbolic contrastive loss such that the prior knowledge can be effectively injected as an inductive bias to influence the learning of topic taxonomy.

Specifically, we set the number of topics at each layer of the hierarchical topic model to be the same as the number of concepts at each layer of the taxonomy, then each topic is assigned a corresponding concept as its semantic prior. Since concepts are connected in the taxonomy, such relations can be transfered accordingly between topics, which provide the basis for picking positive and negative pairs among the topic and word embeddings. Let $\mathcal{T} = \{\boldsymbol{\alpha}^{(l)}\}_{l=0}^{L}$ denote the set of all embeddings, then each embedding $\boldsymbol{\alpha}_i^{(l)} \in \mathbb{R}^D$ is associated with two groups of embeddings as the positive samples and the negative samples, respectively. Then the average hyperbolic contrastive loss is defined as (we omit the superscript l for simplicity of notation)

$$\mathcal{L}_{\text{Contra}} = \mathbb{E}_{\boldsymbol{\alpha}_i \in \mathcal{T}} \left[-\log \frac{\exp\left(\mathcal{S}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^+)/\tau\right)}{\exp\left(\mathcal{S}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^+)/\tau\right) + \sum_{\boldsymbol{\alpha}_i^- \in \mathcal{Q}(\boldsymbol{\alpha}_i)} \exp\left(\mathcal{S}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^-)/\tau\right)} \right] \tag{6}$$

where $\mathcal{S}(\cdot, \cdot)$ is the similarity score function defined in Eq. (5) and τ is the temperature parameter. Note that $\boldsymbol{\alpha}_i^+$ is a positive sample drawn from $\mathcal{P}(\boldsymbol{\alpha}_i)$ and $\mathcal{Q}(\boldsymbol{\alpha}_i)$ is the set of negative samples.

Sampling strategy Inspired by the homophily property (i.e., similar actors tend to associate with each other) in many graph networks [27], we take one-hop neighbors of each anchor, i.e., its parent node and its child nodes as positive samples to maintain the hierarchical semantic information. For the negative samples, we select m embeddings from the non-first-order neighbors that have the highest similarity scores with the anchor embedding.

3.3 NVI Inference

Since most existing NTMs can be viewed as the extensions of the framework of VAEs [13, 14], they generally develop a similar training objective to VAEs, which is to maximize the Evidence Lower BOund (ELBO). For our generative model, the ELBO of each document can be derived as

$$\mathcal{L}_{\text{ELBO}} = -\sum_{l=1}^{L} D_{KL}\left[q(\boldsymbol{\theta}_j^{(l)}|-) \| p(\boldsymbol{\theta}_j^{(l)}|\boldsymbol{\Phi}^{(l+1)}, \boldsymbol{\theta}_j^{(l+1)})\right] + \mathbb{E}_{q(\boldsymbol{\theta}_j^{(1)}|-)}\left[\ln p(\boldsymbol{x}_j|\boldsymbol{\Phi}^{(1)}, \boldsymbol{\theta}_j^{(1)})\right] \tag{7}$$

Algorithm 1 Knowledge Guided Topic Taxonomy Mining

Input: mini-batch size B, number of layers T, adjacent matrix A build from concept taxonomy.
Initialize the variational network parameters $\boldsymbol{\Omega}$ and the word and topic embeddings $\{\boldsymbol{\alpha}^{(l)}\}_{l=0}^{L}$;
while not converge **do**

1. Randomly draw a batch of samples $\{\boldsymbol{x}_j\}_{j=1}^{B}$;
2. Infer variational posteriors for the latent variables of different layers $\{\boldsymbol{\theta}_j^{(l)}\}_{j=1,l=1}^{B,L}$;
3. Derive factor loading matrices $\left\{\boldsymbol{\Phi}^{(l)}\right\}_{l=1}^{L}$ using $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=0}^{L}$ based on Eq. (4);
4. Compute the ELBO on the joint marginal likelihood of $\{\boldsymbol{x}_j\}_{j=1}^{B}$ based on Eq. (7);
5. Compute the hyperbolic contrastive loss using $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=0}^{L}$ and A according to Eq. (6);
6. Update $\boldsymbol{\Omega}$ and $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=0}^{L}$ using gradients $\nabla_{\boldsymbol{\Omega},\boldsymbol{\alpha}^{(l)}}\mathcal{L}\left(\boldsymbol{\Omega},\{\boldsymbol{\alpha}^{(l)}\}_{l=0}^{L};\{\boldsymbol{x}_j\}_{j=1}^{B}\right)$;

end while

where the first term is the Kullback-Leibler divergence that encourages the approximate posterior $q(\boldsymbol{\theta}_j^{(l)}|-)$ to be close to its prior $p(\boldsymbol{\theta}_j^{(l)})$, and the second term is the expected log-likelihood or reconstruction error. Considering that our generative model employs the gamma distributed latent variables, it brings the difficult of reparameterizing a gamma distributed random variable when we design a sampling-based inference network. Therefore, we instead utilize a Weibull distribution to approximate the conditional posterior inspired by Zhang et al. [17], as the analytic KL expression and efficient reparameterization make it easy to estimate the gradient of ELBO with respect to network parameters. The implementation details of our variational encoder is described in Appendix. B.

Furthermore, to inject the prior knowledge to guide the learning of a topic taxonomy, we train the ELBO jointly with a regularization term specified by the proposed contrastive loss in Section 3.2

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \lambda\mathcal{L}_{\text{Contra}} \tag{8}$$

where λ is the hyper-parameter used to control the impact of the regularization term, its detailed effect is investigated in Appendix D. We summarize our complete learning procedure in Algorithm 1.

Table 1: Statistics of the datasets

	Number of docs	Vocabulary size	Total number of words	Categories
20NG	18,846	8,581	1,446,490	20
TMN	32,597	13,368	592,973	7
WIKI	28,472	20,000	3,719,096	N/A
RCV2	804,414	7,282	60,209,009	N/A

4 Experiments

4.1 Experimental setup

Datasets We conduct our experiments on four benchmark data sets with various sizes and document lengths, including *20Newsgroups* **(20NG)** [45], *Tag My News* **(TMN)** [46], *WikiText-103* **(Wiki)** [47], and *Reuters Corpus Volume II* **(RCV2)** [48]. The statistics of these data sets are presented in Table 1. In particular, TMN is a short text corpus with an average of about 20 words per document; 20NG and TMN are the two corpora that associates with document labels. We here do not consider the labels of RCV2 as there are multiple labels for each document.

Baseline models As baselines, we choose several exemplary ones from the state-of-the-art topic models, including: 1) **LDA** [7], one of the most widely-used topic models; 2) **ProdLDA** [16], an NTM which replaces the mixture model in LDA with a product of experts; 3) **ETM** [25], an NTM that marries conventional topic models with word embeddings; 4) **WHAI** [17], a hierarchical NTM which develops a deep Weibull variational encoder based on PGBN [39]; 5) **SawETM** [26], which proposes a Sawtooth Connection module to build the dependencies between topics at different layers;

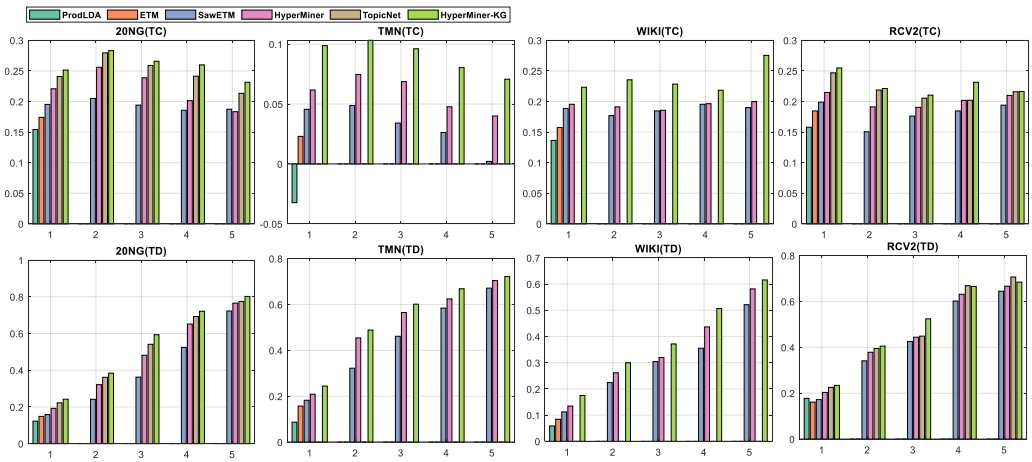

Figure 3: The performance comparison of different models on the topic quality. The top row shows the topic coherence score, i.e., NPMI, and the bottom row displays the topic diversity score. The horizontal axis represents the index of the layers and we set up 5 layers for all the hierarchical topic models. The result of TopicNet on 20NG and TMN is missing because of memory overflow for large vocabulary size.

6) **TopicNet** [29], a knowledge-based hierarchical NTM that guides topic discovery through prior semantic graph. All baselines are implemented meticulously in accordance with their official code.

Evaluation metrics We aim to evaluate the performance of our model in terms of both topic quality and document representation. For topic quality, we adopt topic coherence (**TC**) and topic diversity (**TD**) as performance metrics. Given the reference corpus, TC measures the interpretability of a topic by computing the semantic coherence in the most significant words [49]. Specifically, we apply the widely-used Normalized Pointwise Mutual Information (NPMI) [50] computed over the top 10 words of each topic, with the original document collections of each dataset as the reference corpus. The value of NPMI ranges from -1 to 1, and higher values indicate better interpretability. As the name implies, TD measures how diverse the discovered topics are. Following [25], we define TD to be the percentage of unique words in the top 25 words. TD close to 0 means redundant topics. TD close to 1 reflects more varied topics.

On the other hand, since per-document topic proportions can be viewed as unsupervised document representations, we intend to evaluate the quality of such representations by performing document clustering tasks. Concretely, we report the purity and Normalized Mutual Information (NMI) [51] on two datasets providing the document labels, i.e., 20NG and TMN. With the default training/test split of each dataset, we first train a topic model on the training set, and then use the trained model to extract features $\boldsymbol{\theta}$ of all test documents. Subsequently, we apply the KMeans algorithm on $\boldsymbol{\theta}$ and compute the purity and NMI of the KMeans clusters (denoted by km-Purity and km-NMI). Note that both of the two metrics range from 0 to 1, and higher scores reflect better performance. For the hierarchical topic models, we take the latent units $\boldsymbol{\theta}^{(1)}$ of the first layer as the document feature.

4.2 Experimental results

Topic quality Figure 3 exhibits the comparison results of our model with other baseline models on topic quality. Given that not all discovered topics are interpretable [52], we select the top 50% topics with the highest NPMI and report the average score over those selected topics to evaluate the topic quality comprehensively. The top row of Figure 3 shows that our proposed HyperMiner-KG achieves superior performance at each layer on all datasets, indicating better interpretability. We attribute this improvement to our naturally-designed framework of injecting the tree-structure knowledge in the contrastive manner. In particular, without incorporating the knowledge, HyperMiner outperforms the other deep neural topic models, *i.e.*, ETM and SawETM, at almost all layers, which validates the superiority of hyperbolic geometry in representing hierarchical data. From the bottom row of Figure 3, we can find that our proposed HyperMiner-KG achieves better or at least competitive performances against other methods, leading to more varied topics. By incorporating

Table 2: km-Purity and km-NMI for document clustering. The best and second scores of each dataset are highlighted in boldface and with an underline, respectively. The embedding dimension for embedded topic models is set as 50.

Method	20NG		TMN	
	km-Purity	km-NMI	km-Purity	km-NMI
LDA [7]	38.43 $\pm$ 0.52	35.98 $\pm$ 0.39	48.17 $\pm$ 0.69	30.96 $\pm$ 0.78
ProdLDA [16]	39.21 $\pm$ 0.63	36.52 $\pm$ 0.51	55.28 $\pm$ 0.67	35.57 $\pm$ 0.72
ETM [25]	42.68 $\pm$ 0.71	37.72 $\pm$ 0.64	59.35 $\pm$ 0.74	38.75 $\pm$ 0.86
WHAI [17]	40.89 $\pm$ 0.35	38.90 $\pm$ 0.27	58.06 $\pm$ 0.45	37.34 $\pm$ 0.48
SawETM [26]	43.36 $\pm$ 0.48	41.59 $\pm$ 0.62	61.13 $\pm$ 0.56	40.78 $\pm$ 0.63
TopicNet [29]	42.94 $\pm$ 0.41	40.76 $\pm$ 0.53	60.52 $\pm$ 0.50	40.09 $\pm$ 0.54
HyperETM	43.63 $\pm$ 0.51	39.06 $\pm$ 0.64	61.22 $\pm$ 0.62	40.52 $\pm$ 0.71
HyperMiner	44.37 $\pm$ 0.38	42.83 $\pm$ 0.45	62.96 $\pm$ 0.48	41.93 $\pm$ 0.52
HyperMiner-KG	**45.16** $\pm$ 0.35	**43.65** $\pm$ 0.39	**63.84** $\pm$ 0.43	**42.81** $\pm$ 0.47

Table 3: Accuracy for document classification on 20NG, with different embedding dimensions for embedded topic models.

	$D = 2$	$D = 5$	$D = 10$	$D = 20$	$D = 50$
ETM	19.87 $\pm$ 0.81	33.64 $\pm$ 0.69	39.06 $\pm$ 0.54	42.13 $\pm$ 0.47	43.85 $\pm$ 0.42
HyperETM	**24.33** $\pm$ 0.76	**36.57** $\pm$ 0.65	**40.92** $\pm$ 0.56	**43.04** $\pm$ 0.43	**44.38** $\pm$ 0.40
SawETM	16.74 $\pm$ 0.78	27.05 $\pm$ 0.66	31.68 $\pm$ 0.51	34.06 $\pm$ 0.42	35.42 $\pm$ 0.37
HyperMiner	**20.16** $\pm$ 0.80	**29.73** $\pm$ 0.63	**33.04** $\pm$ 0.49	**34.98** $\pm$ 0.41	**36.01** $\pm$ 0.36
TopicNet	20.29 $\pm$ 0.58	31.26 $\pm$ 0.51	34.57 $\pm$ 0.45	36.84 $\pm$ 0.39	38.02 $\pm$ 0.36
HyperMiner-KG	**22.83** $\pm$ 0.55	**33.15** $\pm$ 0.50	**36.28** $\pm$ 0.43	**38.11** $\pm$ 0.40	**39.46** $\pm$ 0.34

the hierarchical prior knowledge, both TopicNet and HyperMiner-KG result in topics with higher diversity, and HyperMiner-KG performs better. This is mainly because the contrastive loss pushes topic embeddings guided by different semantics far away. Finally, we find that although some baseline methods get higher TD scores than our model at shallow layers, it is difficult for them to achieve a higher NPMI simultaneously.

Document representation Table 2 shows the clustering performance of different models. We run all the models in comparison five times with different random seeds and report the mean and standard deviation. From the results presented above, we have the following remarks: **i)** For all the evaluation metrics, our proposed improved versions perform consistently better than their prototypical models (refer to HyperETM versus ETM, and HyperMiner versus SawETM), which demonstrates that the introduced hyperbolic embeddings are beneficial to both the discovery of high-quality topics and the learning of good document representations. **ii)** As a knowledge-guided topic model, HyperMiner-KG achieves a significant improvement over the base model SawETM, while TopicNet suffers a slight performance degradation compared to SawETM, which also serves as its base model in the original paper. This observation shows that with the hyperbolic contrastive loss, our model can not only inject the knowledge successfully into the learning of hierarchical topics, but also achieve a better balance among the comprehensive metrics of topic modeling. **iii)** The superior performance of our model on TMN also suggests its potential for short text topic modeling.

To further investigate the effectiveness of our method in different dimensional embedding spaces, we proceed to compare the extrinsic predictive performance of document representations through classification tasks. Consistent with the practice in clustering tasks, we first collect the features of training documents $\boldsymbol{\theta}_{tr}$ and test documents $\boldsymbol{\theta}_{te}$ separately, inferred by a well-trained topic model. Afterward, we train an SVM classifier using $\boldsymbol{\theta}_{tr}$ and the corresponding labels. Finally, we use the trained classifier to predict the labels of $\boldsymbol{\theta}_{te}$ and compute the accuracy. Table 3 illustrates the

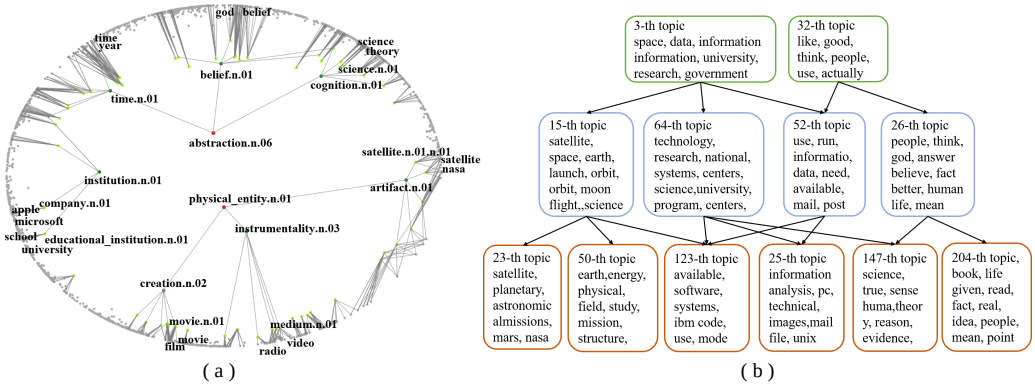

Figure 4: (a) Visualization of 2D hyperbolic embedding space learned by HyperMiner-KG, where colored points represent topic embeddings with the corresponding concepts annotating nearby, grey points denote word embeddings, and the solid line connecting two points comes from the prior taxonomy, indicating they follow the hypernym relation. (b) Illustration of the topic taxonomy learned by a 5-layer HyperMiner-KG, we only show example topics at the middle three layers.

classification results of different embedded topic models. From the table we can see that the improved variants with our method surpass their compared baselines at any dimensionality setting. Especially for the low-dimensional embedding space, there is a significant performance gap between the two, indicating the natural advantage of hyperbolic spaces in modeling hierarchical relations.

Visualization of embedding space and topics As words and topics in our model are embedded into the shared hyperbolic space, with the unique tree-likeness properties of hyperbolic space, we believe that hierarchical structures could be better captured by the distance between embeddings in such a space. This is particularly useful for incorporating prior structural knowledge into hierarchical topic models to mine an interpretable topic taxonomy. To validate this idea, we train a 5-layer HyperMiner-KG on 20NG, with the dimensionality of the embedding space setting as 2. The learned embedding space is visualized in Figure 4(a), from which we can observe that those topic embeddings at the top layers, which are guided by relatively abstract concepts, tend to fall close to the origin of the disk. In contrast, those topics guided by specific concepts have learned their embeddings close to the boundary. More importantly, we find that the distances between learned embeddings reflect the prior hierarchical relations well. For example, the closest topic to the word "university" exactly corresponds to the concept of "educational institution". Moving towards the origin, the closest topic is guided by the concept of "institution". This demonstrates the guidance of prior knowledge plays a key role in learning the continuous embeddings of hierarchies, which should be attributed to our proposed hyperbolic contrastive loss. Moreover, we display the top words of the learned topics in Figure 4(b). From the semantic meaning of each topic and the correlations of topics between different layers, our model does learn an interpretable topic taxonomy.

5 Conclusion

This paper presents a novel framework to improve existing embedded topic models by introducing hyperbolic embeddings to represent words and topics. In using the hyperbolic distance to measure the semantic similarity between two embeddings, the model is encouraged to exploit the underlying semantic hierarchy among the words to find more interpretable topics. In addition, with our designed hyperbolic contrastive loss, the prior structural knowledge is effectively injected into the hierarchical topic models, thus facilitating the learning of a meaningful topic taxonomy. Our method has shown appealing properties that can overcome several shortcomings of existing embedded topic models. Extensive experiments have been conducted, showing that our method achieves consistent performance improvements in both discovering quality topics and deriving useful document representations for regular and short texts.

References

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

[2] David Mimno, Hanna Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889, 2009.

[3] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1):157–208, 2012.

[4] Shuhui Jiang, Xueming Qian, Jialie Shen, Yun Fu, and Tao Mei. Author topic model-based collaborative filtering for personalized poi recommendations. *IEEE transactions on multimedia*, 17(6):907–918, 2015.

[5] Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. Topic compositional neural language model. In *International Conference on Artificial Intelligence and Statistics*, pages 356–365. PMLR, 2018.

[6] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24 (10):2733–2742, 2020.

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

[8] Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16, 2003.

[9] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[10] Jon Mcauliffe and David Blei. Supervised topic models. *Advances in neural information processing systems*, 20, 2007.

[11] Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*, pages 1462–1471. PMLR, 2012.

[12] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):256–270, 2014.

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

[14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

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

[16] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[17] Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. Whai: Weibull hybrid autoencoding inference for deep topic modeling. *arXiv preprint arXiv:1803.01328*, 2018.

[18] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*, 2019.

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

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[21] James Petterson, Wray Buntine, Shravan Narayanamurthy, Tibério Caetano, and Alex Smola. Word features for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 2010.

[22] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313, 2015.

[23] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174, 2016.

[24] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020.

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

[26] Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, pages 2903–2913. PMLR, 2021.

[27] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[28] Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Reducing the rank in relational factorization models by including observable patterns. *Advances in Neural Information Processing Systems*, 27, 2014.

[29] Zhibin Duan, Yi Xu, Bo Chen, Chaojie Wang, Mingyuan Zhou, et al. Topicnet: Semantic graph-guided topic discovery. *Advances in Neural Information Processing Systems*, 34, 2021.

[30] Mikhael Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, 1987.

[31] Matthias Hamann. On the tree-likeness of hyperbolic spaces. In *Mathematical proceedings of the cambridge philosophical society*, volume 164, pages 345–361. Cambridge University Press, 2018.

[32] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018.

[33] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018.

[34] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

[35] Hyunghoon Cho, Benjamin DeMeo, Jian Peng, and Bonnie Berger. Large-margin classification in hyperbolic space. In *The 22nd international conference on artificial intelligence and statistics*, pages 1832–1840. PMLR, 2019.

[36] J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[37] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.

[38] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2018.

[39] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *The Journal of Machine Learning Research*, 17(1):5656–5699, 2016.

[40] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[41] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.

[42] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR, 2020.

[43] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

[44] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*, pages 259–270, 2020.

[45] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

[46] Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. Classification of short texts by deploying topical annotations. In *European Conference on Information Retrieval*, pages 376–387. Springer, 2012.

[47] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[48] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr): 361–397, 2004.

[49] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[50] Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22, 2013.

[51] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[52] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Dirichlet belief networks for topic structure learning. *Advances in neural information processing systems*, 31, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We describe the limitations of our word in the appendix
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the negative societal impacts in the appendix
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They will be included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]