Hierarchical Deep Document Model

Yi Yang, John P. Lalor, Ahmed Abbasi, *Senior Member, IEEE,* Daniel Dajun Zeng, *Fellow, IEEE*

***Abstract*—Topic modeling is a commonly used text analysis tool for discovering latent topics in a text corpus. However, while topics in a text corpus often exhibit a hierarchical structure (e.g., cellphone is a sub-topic of electronics), most topic modeling methods assume a flat topic structure that ignores the hierarchical dependency among topics, or utilize a predefined topic hierarchy. In this work, we present a novel Hierarchical Deep Document Model (HDDM) to learn topic hierarchies using a variational autoencoder framework. We propose a novel objective function, sum of log likelihood, instead of the widely used evidence lower bound, to facilitate the learning of hierarchical latent topic structure. The proposed objective function can directly model and optimize the hierarchical topic-word distributions at all topic levels. We conduct experiments on four real-world text datasets to evaluate the topic modeling capability of the proposed HDDM method compared to state-of-the-art hierarchical topic modeling benchmarks. Experimental results show that HDDM achieves considerable improvement over benchmarks and is capable of learning meaningful topics and topic hierarchies. To further demonstrate the practical utility of HDDM, we apply it to a real-world medical notes dataset for clinical prediction. Experimental results show that HDDM can better summarize topics in medical notes, resulting in more accurate clinical predictions.**

***Index Terms*—Topic Modeling, Deep Learning, Hierarchical Modeling, Hierarchical Neural Topic Model, Textual Analysis.**

I. Introduction

THE rapid growth of unstructured text data has provided an exciting opportunity for researchers and practitioners to arrive at novel insights. With the proliferation of such text data, topic modeling methods that can automatically uncover latent and relevant topics from a set of textual documents (*corpus*) are of great importance. Topic models have been widely used to make sense of textual data in various contexts including social media [1]–[3], e-commerce [4], [5], and personalized recommendation [6]. For example, marketers use topic modeling methods to analyze high volumes of consumer reviews to understand word-of-mouth [7], whereas risk management teams use such models to monitor user-generated content for product defects [8]. Topic models have also been used for knowledge discovery across papers, tools, and datasets appearing in scientific communities [9].

Topic modeling methods usually assume a flat topic structure. That is, all topics have the same level of abstraction and there are no hierarchical relations among topics. For example, when applying topic modeling to a collection of hotel consumer reviews, the model may learn latent topics of *booking reservations*, *customer services*, *booking*, and *hotel location*. However, in practice, topics of words and documents can be naturally organized into hierarchies. A consumer review on the topic of *booking reservations* is also relevant to the more general topics of *booking* and *customer services*. Moreover, topic hierarchies also manifest at the document topic representation level. For example, in writing a clinical note, a physician may describe a mild symptom in general but a critical symptom in considerable detail. It is therefore important to distinguish general topics from specific topics in order to better summarize information from a text corpus. As a result, neural variational inference based methods such as neural topic model (NTM) [10], which represent topics as a flat structure, often lead to suboptimal modeling in certain hierarchical contexts. Compared to flat topic models which suffer from data sparsity and are prone to overfitting [11], [12], a hierarchical approach can combat these shortcomings. For example, if data are too sparse to enable the term *Bronchiolitis* to be placed within the *Bronchitis* concept, the term can instead be appropriately modeled within the topic *Acute respiratory infections*. Moreover, a hierarchical model is particularly suitable for industry-level applications when data is plentiful. For instance, it is reported that Google's Rephil, a core application in Google's online advertising service, uses a hierarchical "noisy-or" network to model topic relations [13].

Prior topic modeling work has proposed several hierarchical topic model extensions based on a Bayesian inference framework. However, the effectiveness of these models in modeling topic hierarchies is usually limited due to their inflexibility in modeling the latent variable relationships (e.g., most rely on a pre-defined topic hierarchy, etc.). We discuss these limitations and gaps in greater detail in the Literature Review Section. In this work, we propose a novel hierarchical deep document model (HDDM) based on the deep variational autoencoder (VAE) framework [14], [15]. The most straight-forward approach to applying VAE to a hierarchical topic model structure is to simply stack multiple neural network layers together in the encoder and decoder. However, simply stacking multiple layers does not necessarily learn good feature hierarchies [16], [17]. To tackle this challenge, we recursively correct the generative distribution with a novel objective function, sum of log likelihood, instead of the conventional evidence lower bound objective (ELBO). Our proposed objective directly models and optimizes the topic-word distribution from all topic levels and can recover the general-specific topic hierarchies.

We evaluate our proposed HDDM method on four real-world text datasets, covering a wide range of applications. Experimental results show that HDDM can significantly improve

This work was partially supported by the U.S. NSF under Grants IIS-2403438, IIS-2039915, CMMI-2240347, and NNSF of China under Grant 72293575.

Y. Yang is with Hong Kong University of Science and Technology, Hong Kong.

J.P. Lalor and A. Abbasi are with University of Notre Dame, Indiana, USA.

D. Zeng is with Institute of Automation, Chinese Academy of Sciences and the University of Chinese Academy of Sciences, Beijing, China.

topic modeling performance, as measured by the model likelihood on unseen data, compared with a battery of state-of-the-art hierarchical topic modeling methods. Moreover, HDDM also enhances topic quality, as evidenced by improvements in both topic coherence and topic diversity. We also visualize the learned topic hierarchies and find that HDDM can discover meaningful and interpretable topic groups.

To further demonstrate the practical value of HDDM, we conduct a clinical prediction task to infer patients' diagnoses from clinical notes. This clinical prediction task can help with treatments, interventions, and healthcare policies. Compared with a battery of state-of-the-art supervised approaches, including deep learning and the bidirectional encoder representations from transformers (BERT) model (Devlin et al. 2018), our unsupervised approach HDDM combined with simple logistic regression substantially improves prediction accuracy.

The main contributions of this work are two-fold. First, we propose a novel topic modeling method, HDDM, that uncovers latent topic hierarchies from text collections. In addition to the proposed hierarchical deep architecture, we develop a novel objective function that facilitates effective learning. Compared to traditional flat structure and hierarchical topic models, HDDM can more effectively learn topic hierarchies directly from the data without a predefined the hierarchy structure. This results in better summarization of the latent semantics in documents to enhance topic analysis and subsequent tasks. Second, we demonstrate in the experiments that HDDM significantly outperforms existing benchmarks in topic modeling. Using a real-world clinical prediction task, we also show that the improvement attributable to HDDM in topic modeling can further lead to more accurate downstream clinical prediction.

Our work has important research implications. By using deep learning and a novel objective function, we allow more accurate topic modeling in the various contexts where domain/testbed-specific hierarchy knowledge is unavailable beforehand and must be learned inductively, in an end-to-end manner. By showing the effectiveness not only for topic modeling but also in a downstream task, our work also has implications for the rising trend of predictive models featurized with topic-model based predictor variables. Furthermore, our work has practical implications for practitioners leveraging text analytics in a broad array of applications where accurate topic analysis may facilitate enhanced decision making [18]. We open-source the implementation of HDDM at https://github.com/yya518/hddm.

The remainder of this paper is organized as follows. In the next section, we review related method development in learning latent topic hierarchies from text corpora. Following the literature review, we present our proposed novel topic modeling approach, HDDM. In the ensuing section, we describe the four real-world text datasets in our testbed and rigorously evaluate our approach against various benchmarks. Then, we examine an application of HDDM in healthcare analytics to infer patient diagnoses from clinical notes. Lastly, we conclude the paper with discussions and implications.

II. Literature Review

Topic modeling is a frequently used tool for discovering latent semantics in a large collection of documents. Latent Dirichlet Allocation (**LDA**) [19] is one of the most influential topic modeling examples. We have witnessed growing popularity in the usage of novel topic modeling methods to tackle task-specific problems [20]–[23]. One common assumption in topic modeling is that the topics have a flat structure. That is, all topics have the same level of abstraction and there are no hierarchical relations among topics. However, in practice, topics can be naturally organized into hierarchies, where a more general concept can include several more-specific topics. Compared to the flat topic model, a hierarchical topic structure can improve document modeling and generalization performance [24]. To this end, several hierarchical topic models based on the statistical Bayesian inference framework have been proposed. Instead of using a flat Dirichlet distribution as the topic distribution prior, these methods adopt a hierarchical topic prior or non-parametric prior to explicitly model the topic hierarchy. For example, Hierarchical LDA (**hLDA**) extends the LDA model by assuming a two-level topic generative process following the Chinese Restaurant Process [25]. A further extension of hLDA is the nested Hierarchical Dirichlet Process (**nHDP**), which is a tree-structured generative model of text that generalizes the nested Chinese Restaurant Process [26]. Both **hLDA** and **nHDP** model the tree structure as a random variable, defined over a flexible (potentially infinite in number) topic space. However, in practice the infinite models are truncated to a maximum size. The Pachinko allocation model (**PAM**) introduces multiple levels of latent supertopics on top of the basic latent topics [24]. Each supertopic is a distribution over the topics at the next level below. In order to efficiently model the hierarchical topic prior and accelerate the inference of the PAM model, prior work proposed a Sparse Backoff Tree prior (**SBT**) [12]. To summarize, Bayesian inference-based hierarchical modeling approaches usually replace LDA's flat Dirichlet prior with complicated hierarchical priors and explicitly define the path between topic and sub-topics.

One drawback of the Bayesian inference-based approaches is that they make strong assumptions on the latent variable distribution. These assumptions limit the generalizability of the models, which can easily lead to low model fitness, and possibly to suboptimal predictive power [10], [27]. Moreover, model inference and parameter estimation often become more complex as topic models grow more expressive. To avoid explicitly defining latent variable distributions, recent neural topic modeling work approximates the intractable distributions over the latent variables with variational autoencoders and thus attains non-linear complex representations for documents with good generalization ability [10], [14], [20], [28]. However, these neural topic models are still flat, and it is challenging to model latent topic hierarchies effectively using deep generative models like deep VAEs for two reasons.

First, a straightforward way to extend a deep VAE model to a hierarchical model is to stack multiple layers of encoder/decoder on top of each other. However, simple layer stacking does not necessarily lead to good latent hierarchies. Hierarchical latent variable models, by stacking multiple layers of encoder/decoder, have trouble learning structured features [17]. Therefore, recent efforts in building deep learning based hierarchical topic models focus on specifying the topic hierarchy distributions [29]–[31]. For example, the tree-structured neural topic model (**TSNTM**) uses a specially designed doubly-recurrent neural network to model the hierarchical topic distribution [30]. TSNTM can be extended to a non-parametric Tree-Structured Neural Topic Model (**nTSNTM**). The assumption on the hierarchical topic distribution restricts the expressive power of topic modeling, and these methods usually achieve unsatisfactory performance in document modeling as measured by goodness-of-fit [31].

Second, as noted above, simple layer stacking does not necessarily lead to good feature hierarchies. Several works have tackled the difficulties of training deep VAEs [16], [32] by proposing to add dependencies to the corresponding layer of encoded documents when modeling the distribution of latent layers. However, in this way, the dependency between different latent layers becomes too complex to be modeled as topic-subtopic distributions. In other words, the evidence lower bound (ELBO), the default choice of training VAE models, is no longer a suitable learning objective because the intermediate latent layers of a hierarchical VAE cannot be parameterized as topic distributions. Forcing the intermediate latent layers to exhibit hierarchical topic distributions requires designing a complicated fixed form of topic parameterization which may restrict the complexity of the topic distribution, thus limiting the modeling capacity. For example, Weibull Hybrid Autoencoding Inference (**WHAI**) models its topic proportions as a series of Gamma distributions with factorized shape parameters depending on the previous layer [33], [34].

Third, as transformer-based language models advance the field of NLP, clustering-based topic modeling methods that utilize contextualized document embeddings as input representations become a viable solution [35]–[38]. In these methods, instead of employing bag-of-words representation as document inputs like traditional and neural topic models, they leverage contextualized document embeddings as inputs. Then, they utilize off-the-shelf clustering methods such as DBSCAN to group documents into clusters, treating each cluster as a unique topic. Subsequently, words with high weights, such as TF-IDF, are considered as representative of topics. Due to the flexibility of these methods, BERTopic can also be used to learn topic hierarchies by employing hierarchical clustering methods [35].

Lastly, another research stream has focused on constructing hierarchical information in topic models using context and/or problem-specific hierarchical information and structure [39]. Examples include distant supervision via citation graphs [40], domain knowledge [9], [41], [42], social media hashtags [43], and social roles [44], not to mention explicitly supervised topic models [4]. For instance, Guo et al [40] leverage citation networks to learn a two-level topic model, where the topics learned from cited documents are the topics from which a document level topic allocation is sampled. Guided Hierarchical Topic Models [41] use a Dirichlet Forest prior and predefined domain knowledge as extra supervision. There has also been interesting work in topic re-estimation [45] which is out of scope for this work. Our method, and our benchmarks, are unsupervised and intended for general-purpose testbeds where hierarchical topic hierarchies may manifest (i.e., not specific to citation network or social media hashtags, etc.), and where inductively modeling them might improve topic modeling and downstream prediction tasks leveraging the generated topic hierarchy.

TABLE I
EXTANT LITERATURE ON HIERARCHICAL TOPIC MODELING. DL=DEEP LEARNING BASED, LTM=LATENT TOPIC MODELING, PTH=PREDEFINED TOPIC HIERARCHY, TMH=TOPIC MODELING HIERARCHY

Model	DL	LTM	PTH	TMH
hLDA [25]	No	Conjugate Prior	Yes	Nested Chinese Restaurant Process
nHDP [26]	No	Conjugate Prior	Yes	Nested Dirichlet Process
PAM [24]	No	Conjugate Prior	Yes	Hierarchical Topic Prior
SBT [12]	No	Conjugate Prior	Yes	Hierarchical Topic Prior
WHAI [33]	Yes	Hierarchical Gamma distribution	No	Multi-level VAE (Directly stacking)
SawETM [34]	Yes	Hierarchical Gamma distribution	No	Multi-level VAE (Directly stacking)
HNTM [29]	Yes	Neural Network	Yes	Multi-level VAE (Directly stacking)
TSNTM [30]	Yes	Neural Network	Yes	Doubly-Recurrent Neural Networks
HDDM (ours)	Yes	Neural Network	No	Multi-level Hierarchial VAE

A. Gaps and Novelty of Proposed Method

Having established the literature on extant hierarchical topic modeling methods, we now describe our method novelty. First, we propose a novel hierarchical deep document model, HDDM, based on the variational autoencoder framework. Instead of simply stacking multiple layers together, we explicitly model the dependencies between layers so that the top layer governs the distribution of the sub-layers, indicating a hierarchical structure. HDDM does not need to pre-specify the topic hierarchy; instead we allow the topic hierarchies (general-specific topic dependency) to emerge from the data itself. Second, to address the challenge of using ELBO as the objective, we propose a novel objective function by summing up the log-likelihood of different latent VAE layers. We prove that this objective is proportional to the model posterior. This objective allows us to effectively and efficiently train a deep hierarchical VAE model with topic distribution parameterization. We conceptually summarize our work in comparison with select relevant hierarchical topic modeling methods in Table I.

III. METHOD: A NEURAL APPROACH TO MODELING TEXT HIERARCHIES

A. Deep Generative Modeling for Text Analytics

We base our approach on the deep variational autoencoder (VAE) framework which has been used in various text categorization studies [46], [47]. Recent neural text modeling studies, namely **Gaussian softmax model (GSM)** [10], have used the VAE approach to model the generative process of documents and employed gradient ascent to maximize the objective function via evidence lower bound (ELBO). Compared with Bayesian inference based statistical topic models such as LDA, the deep VAE approach shows advantages in modeling complex document representations, along with strong generalization ability, as it approximates intractable distributions using a deep neural network [14], [28], [48]. We briefly describe the GSM approach below.

Suppose that we have a document collection D and a vocabulary contains V distinct words $\{w_1, \cdots, w_V\}$. Let $d \in \mathbb{R}^V$ be the bag of words representation of a document in D, and contains N_d word tokens $\{x^1, \cdots, x^{N_d}\}$. The generative process for document $d \in D$ is formulated as follows:

$$\theta \sim \mathcal{G}(\mu_0, \sigma_0^2), \tag{1}$$

$$\pi^n \sim \text{Multi}(\theta_d), \qquad \text{for } n \in [1, N_d] \tag{2}$$

$$x^n \sim \text{Multi}(\beta_{\pi^n}), \qquad \text{for } n \in [1, N_d] \tag{3}$$

Assume there are K topics, GSM proposes that topic distribution θ follows a Gaussian-softmax distribution $\mathcal{G}(\mu_0, \sigma_0^2)$, defined as: $z \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $\theta = \text{softmax}(Wz)$. Here the latent variable z follows a diagonal Gaussian distribution with mean μ_0 and variance σ_0^2, and W is the neural network parameter. π^n is the topic assignment for the n-th word token, $n \in [1, N_d]$. $\boldsymbol{\beta}_{\pi^n} \in \mathbb{R}^V$ represents the topic distribution over words given topic assignment π^n.

A key difference in the generative process between Bayesian text modeling and VAE based text modeling is that the latter uses a Gaussian prior with a neural network to parameterize the topic distribution, while the former has to use a conjugate prior such as the Dirichlet distribution.

To do neural variational inference, an inference network $q(\theta|d)$ is constructed to approximate the posterior $p(\theta|d)$. Specifically, $q(\theta|d)$ is formulated as a diagonal Gaussian distribution: $q(\theta|d) = \mathcal{N}(\mu(d), \sigma^2(d))$. Here, $\mu(d)$ and $\log(\sigma^2(d))$ are functions of d that are modeled by two multilayer perceptron neural networks. Since the posterior distribution is intracable, the Evidence Lower Bound (ELBO) is often used as a surrogate objective function to approximate the true posterior distribution with $q(\theta|d)$, and the objective function is to maximize:

$$\mathcal{J}(\Phi) = \mathbb{E}_{q(\theta|d)}[\log p(d|\theta)] - \text{KL}(q(z|d)||p(z)). \tag{4}$$

where Φ denotes the overall model parameters in the inference and generative network. The network architecture of the VAE based text model is presented in Figure 1.

While the deep generative framework benefits from an approximate complex distribution without relying on a conjugate prior, the learned latent variables are flat, such as in GSM; that is, the topics are not grouped hierarchically. Given concept hierarchies are an important knowledge representation in web and text mining [49], [50], we develop a novel deep generative model that reduces the dimensionality of text documents to a hierarchical topic space.

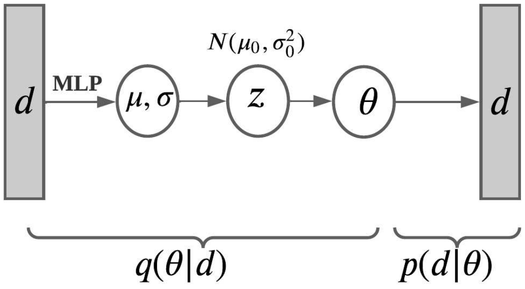


Fig. 1. Variational autoencoder based text modeling.

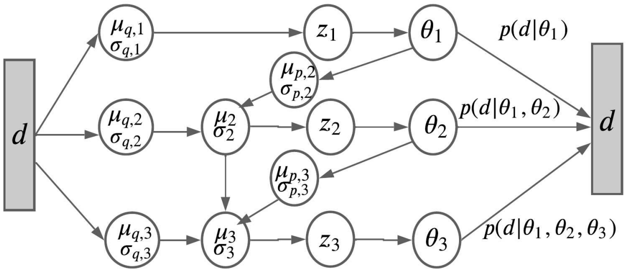


Fig. 2. Our proposed hierarchical topic modeling approach.

B. Learning Topic Hierarchies with a Deep Generative Model

Here we present our hierarchical deep document model, HDDM, which groups latent topics in a text corpus into a topic hierarchy. This model arranges the topics into a tree-structure, with the purpose that more general topics should appear near the root and more specialized topics should appear near the bottom. The key novelty in desiging the hierarchical structure in HDDM is that it allows the latent variable distributions of previous layers to govern the distributions of subsequent layers. That is, the distribution of fine-grained topics is stochastically determined by more coarse-grained topics, allowing latent topic hierarchies to emerge from the data. The HDDM network architecture is shown in Figure 2.

We propose that a document collection can be grouped in a latent topic hierarchy, such that higher-level topics are more coarse-grained and lower-level topics are more fine-grained. Specifically, suppose that we have $L+1$ levels of topics, where θ_0 is fixed as the root topic, θ_1 is the first-level (general) topic vector, and θ_L is the last-level (specific) topic vector. The l-th level has K_l topics and $K_l < K_{l+1}$ for $l \in [0, L]$. We propose that a document can be generated from each of the L layers of topics respectively. That is, a document is a summarization of general topics and specific topics. The generative process of HDDM for document $d \in D$ is described as follows.

$$\theta_l \sim \mathcal{G}(\mu_l, \sigma_l^2), \qquad \text{for } l \in [1, L] \tag{5}$$

$$\pi_l^n \sim \text{Multi}(\theta_l), \qquad \text{for } n \in [1, N_d] \tag{6}$$

$$x^n \sim \text{Multi}(\beta_{\pi_l^n}), \qquad \text{for } n \in [1, N_d] \tag{7}$$

In this generative process, we first let the root topic $\theta_0 = 1$ to be the same for all documents, and assign $\pi_0^n = root$ to every token in every document. Then, the level-l topic distribution θ_l follows a Gaussian-softmax distribution with mean μ_l and variance σ_l^2. Different from GSM where μ_0 and σ_0^2 are hyperparameters such as $\mu_0 = 0$ and $\sigma_0^2 = I$, we define μ_l and σ_l^2 as a function of θ_{l-1}. This means that the prior of θ_l depends on θ_{l-1}.

For the inference network, we design the multi-level VAE such that the latent variable distributions consist of the previous level' latent variables as well as the corresponding level of encoded input. To facilitate topic modeling, we introduce a topic–word distribution parameter matrix $\beta_l \in \mathbb{R}^{K_l \times V}$ for each layer l. We apply the softmax function to ensure that each entry in the column for β_l adds up to 1. For any $l \in [1, L]$, the inference network is defined recursively as follows:

$$q(\theta_l|\theta_{l-1}, d) = \mathcal{N}(W_l\mu_l, (W_l\sigma_l)^2) \tag{8}$$

Here, since the latent variable consists of previous level's latent variable (i.e., the prior of θ_l depends on θ_{l-1}) and the corresponding level's encoded input (i.e., the prior of θ_l also depends on the bag-of-words representation d), we use the inverse-variance weighting as:

$$\mu_l = (\mu_{q,l}\sigma_{q,l}^{-2} + \mu_{p,l}\sigma_{p,l}^{-2})/(\sigma_{q,l}^{-2} + \sigma_{p,l}^{-2}) \tag{9}$$

$$\sigma_l = 1/(\sigma_{q,l}^{-2} + \sigma_{p,l}^{-2}) \tag{10}$$

where $\mu_{p,l}, \sigma_{p,l}^2$ and $\mu_{q,l}, \sigma_{q,l}^2$ are obtained from the neural network as follows:

$$\mu_{p,l} = W_{\mu_{p,l}}\theta_{l-1}, \quad \log(\sigma_{p,l}^2) = \text{softmax}(W_{\sigma_{p,l}}\theta_{l-1}), \tag{11}$$

$$\mu_{q,l} = W_{\mu_{q,l}}d, \quad \log(\sigma_{q,l}^2) = \text{softmax}(W_{\sigma_{q,l}}d). \tag{12}$$

For $l = 1$, as θ_0 is fixed to the root topic, we directly let $\mu_l = \mu_{q,l}, \sigma_l = \sigma_{q,l}$. Then $q(z_l|z_0, z_1, ..., z_{l-1}, d) = \mathcal{N}(\mu_l, \sigma_l)$ and we generate z_l from this distribution.

C. Model Inference and Parameter Estimation

We now discuss the model inference and parameter estimation for HDDM. Traditionally, evidence lower bound (ELBO) is the objective function of deep variational autoencoder modeling (c.f. Eq. 4). However, in HDDM, the intermediate latent layers cannot be parameterized as topic distributions. To force the intermediate latent layers to exhibit topic distributions (leading to a hierarchical topic model), one would have to design a complicated and fixed form of topic parameterization, which might restrict the complexity of the topic distribution and thus limit the modeling capacity. Therefore, we develop a novel objective function, the sum of the log likelihood, to directly model and optimize the topic–word distribution at all topic levels and then approximate the supertopic–subtopic distributions from the topic–word distributions. We define the sum of the log likelihood objective function for document d with N_d word tokens as follows:

$$\mathcal{J}(\Phi) = \frac{1}{L}\sum_{n=1}^{N_d}\sum_{l=1}^{L}\log p(x^n|\theta_l) \tag{13}$$

We show in Theorem 1 that the sum of the log likelihood is proportional to the parameter posterior. Therefore, instead of maximizing ELBO to approximate the posterior distribution, we can directly maximize the sum of the log likelihood.

Theorem 1. $\mathcal{J}(\Phi)$ is proportional to the parameter posterior.

Proof. We start by expanding $p(x^n|\Phi, \theta_0, \theta_1, \dots, \theta_L)$ using Bayes's rule as follows: (n, Φ are omitted for simplicity)

$$p(x|\theta_0, ..., \theta_L) = \frac{p(\theta_0, ..., \theta_L|x)p(x)}{p(\theta_0, ..., \theta_L)} \tag{14}$$

$$= \frac{p(\theta_L|x, \theta_0, ..., \theta_L)...p(\theta_0|x)p(x)}{p(\theta_0, ..., \theta_L)} \tag{15}$$

$$= \frac{p(x|\theta_0, ..., \theta_L)p(\theta_L)}{p(x)} \cdots \frac{p(x|\theta_0)p(\theta_0)}{p(x)} \frac{p(x)}{p(\theta_0, ..., \theta_L)} \tag{16}$$

$$= \frac{\prod_{l=0}^{L} p(x|\theta_0, \dots, \theta_l)}{p(x)^L} \frac{\prod_{l=0}^{L} p(\theta_l)}{p(\theta_0, ..., \theta_L)} \tag{17}$$

$$= \frac{p(x|\theta_0, \theta_1, ..., \theta_L)\prod_{l=1}^{L} p(x|\theta_0, \dots, \theta_l)}{p(x)^L} \frac{\prod_{l=0}^{L} p(\theta_l)}{p(\theta_0, ..., \theta_L)} \tag{18}$$

We then cancel $p(x|\theta_0, \theta_1, ..., \theta_L)$ on both left-hand side and right-hand side, and also $p(\theta_0), ..., p(\theta_L), p(\theta_0, \theta_1, ..., \theta_L)$ are untrainable priors. Then we obtain

$$p(x) = [\prod_{l=1}^{L} p(x|\theta_0, \dots, \theta_l)\frac{\prod_{l=0}^{L} p(\theta_l)}{p(\theta_0, ..., \theta_L)}]^{1/L} \tag{19}$$

$$\propto [\prod_{l=1}^{L} p(x|\theta_0, ..., \theta_l)]^{1/L} \approx [\prod_{l=1}^{L} p(x|\theta_l)]^{1/L} \tag{20}$$

since $p(x|\theta_0, ..., \theta_l) \approx p(x|\theta_l)$. This means that we can approximate the posterior by the geometric mean of likelihood with respect to each topic level. After taking logarithms and summing over the N_d word tokens of the current document d, we obtain the learning objective as

$$\mathcal{J}(\Theta) = \sum_{n=1}^{N_d}\log p(x^n) = \frac{1}{L}\sum_{n=1}^{N_d}\sum_{l=1}^{L}\log p(x^n|\theta_l) \tag{21}$$

This completes the proof. ■

The data likelihood $p(x^n|\theta_l)$ in $\mathcal{J}(\Theta)$ can be obtained by marginalizing out the l-th level topic variable π_l^n:

$$p(x^n|\theta_l) = \sum_{\pi_l^n=1}^{K_l} p(x^n|\pi_l^n)p(\pi_l^n|\theta_l) = \beta_{l,x^n}^T\theta_l, \tag{22}$$

where β_{l,x^n}^T is the transpose of the x^n-th column of the topic-word distribution matrix of the l-th topic level.

D. Topic Hierarchy and Topic-word Re-Ranking

As shown in Figure 2, we explicitly model the multi-level latent variable structure and allow it to emerge from the data. The conditional distribution of the l-th level topic i given the $(l-1)$-th level topic j can be written as:

$$\begin{aligned}
&p(\pi_l = i|\pi_{l-1} = j) = \\
&\quad \frac{1}{p(\pi_{l-1} = j)} \\
&\quad \times \sum_{k=1}^{V} \int_{\theta_{l-1}} [p(\pi_{l-1} = j|x = w_k, \theta_{l-1}) \\
&\quad \times p(\pi_l = i|x = w_k, \theta_{l-1}) p(\theta_{l-1})] d\theta_{l-1},
\end{aligned} \tag{23}$$

where π_{l-1} and π_l are conditionally independent given x and θ_{l-1}. As the priors of π_l are close to uniform and the prior of θ_{l-1} is concentrated in a very small interval, $p(\pi_l = i|\pi_{l-1} = j)$ can be approximated by $\sum_{k=1}^{V}[p(\pi_{l-1} = j|x = w_k)p(\pi_l = i|x = w_k)]$. Next, by further assuming that the prior of a single word x is uniform over the vocabulary, the previous approximation is proportional to

$$\sum_{k=1}^{V}[p(x = w_k|\pi_{l-1} = j)p(x = w_k|\pi_l = i)], \tag{24}$$

which equals to $\beta_{l-1,j}\beta_{l,i}^T$, where $\beta_{l,i}$ is the l-th row of β_i. Finally, we have

$$p(\pi_l = i|\pi_{l-1} = j) \propto \beta_{l-1,j}\beta_{l,i}^T \tag{25}$$

To obtain a deterministic topic hierarchy, we regard the $(l-1)$-th level topic with the highest probability of generating topic i on the l-th level as its parent topic: $\mathrm{argmax}_j\, \beta_{l-1,j}\beta_{l,i}^T$.

Additionally, recent literature has shown that simply re-ranking the topic words as a post-processing technique can enhance the interpretability of the learned topics for probabilistic-based topic models [51]–[53] and for clustering-based topic models [35], [54]. Motivated by this, we re-rank the topic distribution over words as follows:

$$\hat{\beta}_{k,i,l} = \frac{\beta_{k,i,l}}{\sum_{k=1}^{K}\beta_{k,i,l}} \tag{26}$$

This means the re-ranked probability of word i for topic k at level l is normalized by the sum of topic-word probabilities across all topics at level l.

To summarize, HDDM reduces unstructured, high-dimensional text to a hierarchical latent topic space. We make two fundamental contributions. First, we propose a novel deep learning network that explicitly models the hierarchy of latent topic variables. Second, we propose a novel objective function to facilitate the effective learning of hierarchical latent variables.

IV. Dataset and Benchmarks

A. Datasets

We now describe the four real-world text datasets used in our experiments. The four datasets represent a wide range of applications that are of high relevance to research and practice. The dataset statistics are shown in Table II.

- Clinical Notes: We use the MIMIC-III dataset [55]. This dataset contains de-identified electronic health record data from various ICUs at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. It contains 46,520 unique patient hospital admission records and information on 61,532 ICU stays between 2001 and 2012, and thus encompasses a diverse and very large population of ICU patients. The dataset comprises clinical notes made during the patients' time in hospital, including electrocardiography (ECG) reports, radiology reports, physicians' and nurses' notes, and discharge summaries. For each note, the date and time are recorded. We use the data and perform our experiments in accordance with the MIMIC-III data user agreement. The dataset is a benchmark dataset that is widely used in the science community for healthcare studies [56]. Following prior work, we exclude records of hospital admissions with multiple ICU stays or transfers between ICU units. We also exclude records of hospital admissions for patients younger than 18 [56]. The final cohort in our analysis contains 33,798 unique patients with a total of 42,276 ICU stays. We choose 100,000 physician notes as training data and the remaining 41,624 physician notes as a held-out dataset. We use 10,000 physician notes from the 100,000 training set as a validation set.
- Online Reviews: The online review dataset is a collection of movie reviews [57]. It contains 5,006 movie reviews with an average length of 188 words.[1]
- Company Descriptions: We use a text dataset that contains company descriptions. This dataset was included to represent contexts where topic models are applied by researchers and analysts to gauge the information content in firm-related documents and reports. This dataset contains 29,867 company descriptions collected from corresponding Wikipedia company pages [58].
- Newsgroup Posts: The 20 newsgroups dataset is a publicly available[2] and widely used benchmark dataset for evaluating topic modeling performance [59]. The dataset contains 18,670 newsgroup posts on different topics such as sports, politics and computers.

For the Newsgroup Posts data, we adopt the vocabulary provided by Srivastava and Sutton [60] for direct comparison. For the other three datasets, we choose the most frequent 2,000 words (exclude stop words) as the vocabulary for all datasets, because choice of a very large vocabulary is likely to result in slow inference and poor topic generation performance [61].

TABLE II
Dataset statistic descriptions.

Dataset	# Train	# Test	Avg. # Words
Clinical Notes	100,000	41,624	1,460.3
Online Reviews	3,337	1,669	131.6
Company Descriptions	30,169	12,930	42.2
Newsgroup Posts	11,218	7,452	84.6

B. Baselines

We consider the following methods as benchmarks for comparison.

[1] https://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] http://qwone.com/ jason/20Newsgroups/

- **Latent Dirichlet allocation (LDA)** [19] is the most cited and widely used topic modeling method. We adopt the online LDA implementation in Gensim [62].
- **Gaussian softmax model (GSM)** [10] is the neural topic model based on the deep variational autoencoder framework. Similar to LDA, it assumes a flat topic structure.
- **Hierarchical LDA (hLDA)** [25] is the hierarchical extension of LDA. It models a two-level hierarchical topic model and assumes the topic generative process using a nested Chinese Restaurant Process. We use the hLDA paper's originally released C code to implement hLDA.
- **Weibull hybrid autoencoding inference (WHAI)** [33] is a deep learning based hierarchical topic model. It uses a hierarchy of Gamma distributions as the generative network and a Weibull distribution in the inference network. We implement WHAI using the authors' released code.
- **Hierarchical Neural Topic Model (HNTM)** [29] learns a hierarchical topic structure. It explicitly models the dependency between different layers. The evidence lower bound (ELBO) is used as the model objective.
- **Tree-structured Neural Topic Model (TSNTM)** [30] models the topic hierarchies using a tree-structure, with ELBO also used as the model objective. Similarly, [31] propose an non-parametric version of TSNTM (nTSNTM) which assumes an infinite number of topics. For this benchmark, we use the TSNTM model.
- **Sawtooth Factorial Topic Embeddings Guided GBN (SawETM)** [34] captures the dependencies and semantic similarities between the topics in the embedding space and it uses the Gamma belief network for modeling the topic distributions of intermediate layers. We implement SawETM using the authors' released code.

In addition to the probabilistic and neural topic models mentioned above, we also incorporate a clustering-based topic modeling approach **BERTopic** [35]. BERTopic applies a clustering method to contextualized document embeddings and utilizes a reranking technique [51], [63] to identify representative words of topics. In our experiment, we evaluate two versions of BERTopic: **SBERT** and **Ada**. For SBERT, contextualized document embeddings are obtained using the sentence-bert model [64], specifically the `all-MiniLM-L6-v2` model. For Ada, we utilize OpenAI's embedding model `text-embedding-ada-002`. We conduct the experiments using the implementation provided by BERTopic [3].

Among these benchmarks, LDA and GSM are flat topic modeling benchmarks, and HDP, WHAI, TSNTM and HNTM are the hierarchical topic modeling benchmarks. SBERT and Ada are clustering-based topic modeling benchmarks.

V. Document Modeling Experiments

The most common evaluation for a probabilistic topic model is to measure the goodness of fit on held-out test data. In this section, we evaluate HDDM's document modeling capability against the benchmarks.

A. Model Fitness

We measure the log-likelihood of models on the hold-out set using perplexity. Perplexity is computed as a function of the data log-likelihood of a held-out test set:

$$\text{perplexity} = \exp(-\sum_{m=1}^{D_{test}} \frac{1}{N_m} \sum_{n=1}^{N_m} \log p(x^{(n)}|t)) \qquad (27)$$

where D_{test} is the number of testing documents, N_m denotes the number of words in the m-th testing document, and $\log p(x^{(n)}|t)$ is the log likelihood of the n-th word in the m-th document. A lower perplexity number indicates a higher data likelihood, which suggests that the topic model is easier to generalize and better at predicting new unseen documents.

Since the baseline approaches include both non-hierarchical and hierarchical methods, to have a fair comparison, we choose three settings with different number of lowest-level topics: 32, 64 and 128. That is, for LDA and GSM, which are non-hierarchical models, the number of topics is set to 32, 64 or 128. For the hierarchical methods, we set a three-level topic hierarchy with 8-16-32, 16-32-64 and 32-64-128 respectively. The hLDA model infers the topic hierarchies from data and we find that the inferred topic number is consistent across several runs. We also tuned the topic hierarchy from three-levels to two-levels and increased/decreased the intermediate-layers topics (keeping the lowest-level of topics the same). The baseline models' likelihoods do not change much, which is consistent with findings reported in the literature [25], [33].

The main result is presented in Table III. First, we can see that the proposed HDDM approach consistently outperforms the baseline approaches in terms of document modeling performance. This superiority holds true across various datasets and regardless of the number of topics considered. For example, on the Clinical Notes dataset, HDDM achieves a perplexity score of 618.3, which is substantially lower than the benchmarking models. Compared to hLDA, the difference is significant ($p < 0.01$). Second, it outperforms its non-hierarchical, ELBO-based VAE counterpart, GSM, suggesting that more effective hierarchical models can result in better document modeling. Moreover, existing hierarchical neural topic models, including WHAI, SawETM, HNTM and TSNTM, primarily relying on the ELBO training objective, experience a degradation in perplexity. This decline might be largely attributed to the requirement of parametrizing intermediate latent layers as topic distributions in ELBO-based models, potentially limiting the model's capacity. Clustering-based BERTopic approaches SBERT and Ada do not rely on probabilistic models. For instance, BERTopic assumes that documents within the same cluster share the same topic, meaning each document is assigned to one topic. Therefore, we do not report perplexity for SBERT and Ada.

To better understand the impact of HDDM's hierarchical structure on model perplexity, we perform an ablation analysis by varying the topic hierarchies. The effect of the number of hierarchy layers on HDDM's performance is depicted in Table IV. Generally, the hierarchical structured topic model is better than the single-layer topic model in terms of perplexity. For the Newsgroups, Clinical Notes, and Reviews data, per-

[3] https://maartengr.github.io/BERTopic/index.html

TABLE III
PERPLEXITY OF TOPIC MODELS ON FOUR DATASETS. WE REPORT STATISTICAL SIGNIFICANCE COMPARED TO hLDA UNDER A ONE-TAILED T-TEST. $*$ INDICATES $p < 0.01$.

	Clinical Notes			Reviews			Company Descriptions			Newsgroups		
	K=32	K=64	K=128	K=32	K=64	K=128	K=32	K=64	K=128	K=32	K=64	K=128
LDA	964.1	785.9	774.6	1325.0	1318.1	1105.4	813.2	732.4	667.7	1263.4	1084.8	1142.4
GSM	772.3	712.0	746.8	954.3	911.5	$\mathbf{923.1}^*$	547.5	655.7	640.7	871.9	861.2	880.1
hLDA	861.5	745.3	833.5	963.7	1034.7	948.4	697.0	677.6	538.2	1069.9	958.7	921.0
WHAI	864.8	835.9	841.8	984.3	1149.5	962.4	879.3	820.4	773.6	990.4	913.3	893.1
SawETM	852.4	766.9	702.1	943.7	859.2	938.9	655.2	589.6	520.5	930.8	906.6	842.3
HNTM	907.6	822.4	779.2	898.3	1010.2	971.0	752.9	693.3	534.9	1014.1	955.4	879.8
TSNTM	843.2	779.8	754.5	1275.3	1185.6	1037.6	723.4	674.1	665.2	945.8	997.3	912.0
HDDM (ours)	$\mathbf{745.0}^*$	$\mathbf{618.3}^*$	$\mathbf{598.6}^*$	$\mathbf{874.6}^*$	$\mathbf{832.9}^*$	926.1	$\mathbf{310.0}^*$	$\mathbf{267.9}^*$	$\mathbf{233.1}^*$	$\mathbf{675.6}^*$	$\mathbf{639.5}^*$	$\mathbf{684.2}^*$

plexity decreases as the number of topic layers increases up to three levels. In the case of Clinical Notes, the decreasing perplexity trend continues for 4 levels (see 8-16-32-64 row in Table IV). For the Company Descriptions dataset, the trend is not as clear beyond the second level – perplexity increases for the 3 and 4 level hierarchices. This could be because of the nature of the datasets, and the exact choice of topics per level. Nevertheless, the hierarchy ablation shows the value of the hierarchical component of HDMM, which works in conjunction with the sum of log-likelihood objective function to garner enhanced performance relative to existing hierarchical benchmarks.

TABLE IV
PERPLEXITY OF HDDM WITH DIFFERENT TOPIC HIERARCHIES. 16-64 MEANS THAT THERE ARE TWO LEVELS OF TOPICS WITH 16 SUPERTOPICS AND 64 SUBTOPICS.

# Topics	Clinical Notes	Reviews	Company Descriptions	Newsgroups
64	922.2	931.0	294.7	671.0
16-64	618.3	832.9	267.9	639.5
16-32-64	540.9	828.2	268.1	638.8
8-16-32-64	549.8	830.2	271.8	612.1

B. Topic Coherence and Topic Diversity

Examining the learned topic quality is also crucial to evaluate the learned topic model [65], [66]. In this section, we conduct experiments to evaluate HDDM's ability to learn high-quality topic words by considering two metrics: **Topic Coherence** and **Topic Diversity**. Topic coherence measures the interpretability of a topic by measuring the degree of semantic similarity between high probability words in the topic. Specifically, we use the C_V coherence measure, which averages the Normalized Pointwise Mutual Information (NPMI) for every pair of words within a sliding window and returns the mean of the NPMI for the given top words. In a systematic study, this coherence measure gives the largest correlation with human ratings [67]. Specifically, topic coherence C_V for topic k and its top T words $(w_1, w_2, \cdots, w_T)$ is defined as

$$C_V(k) = \frac{1}{T}\sum_{i=1}^{T} \cos(\overrightarrow{\text{NPMI}}(w_i), \overrightarrow{\text{NPMI}}(w_{1:T}))$$

Here $\cos(\cdot)$ is the cosine similarity, and

$$\overrightarrow{\text{NPMI}}(w_i) = \{\text{NPMI}(w_i, w_j)\}_{j=1,\ldots,T}$$
$$\overrightarrow{\text{NPMI}}(w_{1:T}) = \left\{\sum_{i=1}^{T} \text{NPMI}(w_i, w_j)\right\}_{j=1,\ldots,T}$$

where $\text{NPMI}(w_i, w_j)$ is the normalized pointwise mutual information for a pair of words w_i and w_j.

Topic diversity (**TD**) is the percentage of unique words in the top T words of all topics [37], [68]. A small value close to zero for topic diversity indicates that the top T words are duplicated across topics, suggesting redundant topics. It is worth noting that topic coherence and topic diversity should be considered together to measure topic quality because a topic model with high topic coherence might simply repeat coherent words across topics. Therefore, following the approach outlined in [37], we define the overall metric Topic Quality (**TQ**) for the quality of a topic model's topics as the product of its topic coherence and topic diversity.

Additionally, we apply the re-ranking technique (i.e., Equation 26) to all topic modeling baselines, and report the topic coherence, topic diversity, and topic quality scores at the bottom level, i.e., the finest-grained topics, so that the results are comparable [51], [53], [54], [63]. For the LDA baseline, we also consider an alternative without using the topic-word re-ranking technique, denoted as LDA-NR, for benchmarking.

The experiment results for topic coherence, topic diversity, and overall topic quality are presented in Table V. We observe that our method, HDDM, achieves the highest topic quality score (TQ) across most cases within the four different datasets, indicating its capability to generate highly interpretable and diverse topics. In addition, it's noteworthy that the superior performance in TQ is primarily attributed to high topic diversity scores. In several instances, the topic diversity score (TD) of HDDM exceeds 0.90, and sometimes even reaches 1.00, implying its ability in learning distinguishable and non-duplicated topics. This characteristic can further facilitate effective content analysis [66]. It is also worth noting that HDDM consistently outperforms the clustering-based topic model BERTopic (SBERT and Ada) in the learned topic coherence and topic diversity metrics. While BERTopic utilizes pretrained large language models to obtain document embeddings, it regards all documents whose document embeddings belong to a single cluster as one single document and explicitly

TABLE V

TOPIC COHERENCE (C_V), TOPIC DIVERSITY (TD), AND TOPIC QUALITY (TQ) METRICS ON FOUR DATASETS. FOR THE HIERARCHICAL TOPIC MODEL METHODS, INCLUDING HDDM, THE EVALUATION UTILIZES THE BOTTOM-LEVEL (FINEST-GRAINED) TOPIC WORDS. SBERT AND ADA IMPLICITLY RERANK TOPIC-WORDS. ALL OTHER BASELINES IMPLEMENT A POST-PROCESSING TOPIC-WORD RERANKING METHOD [51].

	Clinical Notes									Reviews								
	K=32			K=64			K=128			K=32			K=64			K=128		
	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ
LDA-NR	0.37	0.42	0.16	0.35	0.48	0.17	0.33	0.48	0.16	0.34	0.36	0.12	0.32	0.37	0.12	0.3	0.3	0.09
LDA	0.50	0.71	0.36	0.50	0.47	0.23	0.54	0.41	0.22	0.52	0.86	**0.45**	0.54	0.74	0.40	0.58	0.45	0.26
GSM	0.38	0.54	0.20	0.40	0.57	0.23	0.33	0.56	0.19	0.42	0.72	0.30	0.42	0.45	0.19	0.45	0.30	0.13
hLDA	0.50	0.55	0.27	0.46	0.63	0.29	0.51	0.68	0.35	0.49	0.87	0.42	0.39	0.65	0.25	0.36	0.68	0.25
WHAI	0.51	0.48	0.24	0.48	0.74	0.35	0.55	0.61	0.33	0.44	0.45	0.20	0.36	0.81	0.30	0.47	0.56	0.26
SawETM	0.47	0.68	0.32	0.55	0.47	0.26	0.57	0.75	0.43	0.50	0.76	0.38	0.55	0.70	0.39	0.41	0.66	0.27
HNTM	0.44	0.72	0.31	0.55	0.52	0.28	0.50	0.67	0.33	0.36	0.73	0.26	0.47	0.73	0.35	0.33	0.65	0.22
TSNTM	0.54	0.64	0.34	0.57	0.48	0.27	0.49	0.57	0.28	0.53	0.78	0.41	0.52	0.75	0.39	0.47	0.57	0.27
SBERT	0.47	0.68	0.32	0.41	0.52	0.22	0.48	0.86	0.41	0.36	0.65	0.24	0.38	0.61	0.23	0.37	0.58	0.21
Ada	0.42	0.75	0.32	0.50	0.79	**0.39**	0.52	0.83	0.43	0.41	0.68	0.28	0.39	0.63	0.24	0.40	0.62	0.25
HDDM	0.48	0.87	**0.42**	0.42	0.81	0.34	0.53	0.95	**0.50**	0.43	0.98	0.43	0.49	0.86	**0.42**	0.54	0.58	**0.31**

	Company Descriptions									Newsgroups								
	K=32			K=64			K=128			K=32			K=64			K=128		
	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ	C_V	TD	TQ
LDA-NR	0.44	0.52	0.23	0.45	0.52	0.23	0.41	0.53	0.22	0.49	0.54	0.26	0.46	0.49	0.23	0.45	0.54	0.24
LDA	0.62	0.81	**0.50**	0.65	0.68	0.44	0.67	0.58	0.39	0.34	0.89	0.30	0.52	0.74	0.38	0.54	0.58	0.31
GSM	0.53	0.85	0.45	0.56	0.82	0.45	0.40	0.46	0.19	0.51	0.84	0.43	0.41	0.95	0.38	0.40	0.68	0.27
hLDA	0.48	0.70	0.33	0.36	0.73	0.26	0.56	0.67	0.38	0.42	0.68	0.29	0.47	0.67	0.32	0.45	0.65	0.29
WHAI	0.43	0.74	0.32	0.48	0.71	0.35	0.52	0.55	0.29	0.53	0.73	0.39	0.50	0.74	0.37	0.54	0.58	0.31
SawETM	0.49	0.90	0.44	0.49	0.77	0.38	0.61	0.57	0.35	0.59	0.74	0.44	0.64	0.49	0.31	0.58	0.59	0.34
HNTM	0.54	0.80	0.43	0.50	0.63	0.32	0.61	0.37	0.22	0.48	0.77	0.37	0.41	0.81	0.34	0.39	0.73	0.28
TSNTM	0.47	0.63	0.30	0.34	0.89	0.30	0.29	0.43	0.12	0.37	0.87	0.32	0.50	0.59	0.30	0.42	0.86	0.36
SBERT	0.51	0.85	0.43	0.48	0.72	0.35	0.46	0.63	0.29	0.47	0.78	0.37	0.49	0.69	0.34	0.46	0.57	0.26
Ada	0.56	0.63	0.35	0.57	0.76	0.43	0.58	0.85	**0.49**	0.48	0.83	0.40	0.52	0.75	0.39	0.55	0.60	0.33
HDDM	0.50	1.00	**0.50**	0.48	0.99	**0.47**	0.48	0.91	0.44	0.50	0.93	**0.46**	0.47	0.98	**0.46**	0.43	0.99	**0.43**

assumes that a document only exhibits one topic. In practice, a document can naturally relate to multiple topics. Moreover, BERTopic uses a simple TF-IDF-based technique to find the most prominent words within a cluster (topic). In contrast, the proposed HDDM models leverage deep neural networks and explicitly model the hierarchical relationships between topics. Even though the input document is represented in a bag-of-words fashion, HDDM can still effectively discover the hierarchical and topical relationships between words, thereby achieving higher topic coherence and diversity.

To better understand the topic quality of our method, we further evaluate the topic coherence and topic diversity at different hierarchical levels. Specifically, we set the overall hierarchy to three levels and evaluate different hierarchical settings: 8-16-32, 16-32-64, and 32-64-128 respectively. We measure the average topic coherence, topic diversity, and topic quality across the three hierarchical settings and across the four datasets. The results are presented in Table VI. It shows that HDDM consistently achieves high topic coherence, topic diversity, and corresponding topic quality at all three levels. Moreover, the topic quality achieves the highest at the top level, indicating that HDDM is capable of learning distinguishable and diverse coarse-grained topics. The topic quality gradually and slightly decreases as the hierarchy goes down from coarse-grained to fine-grained.

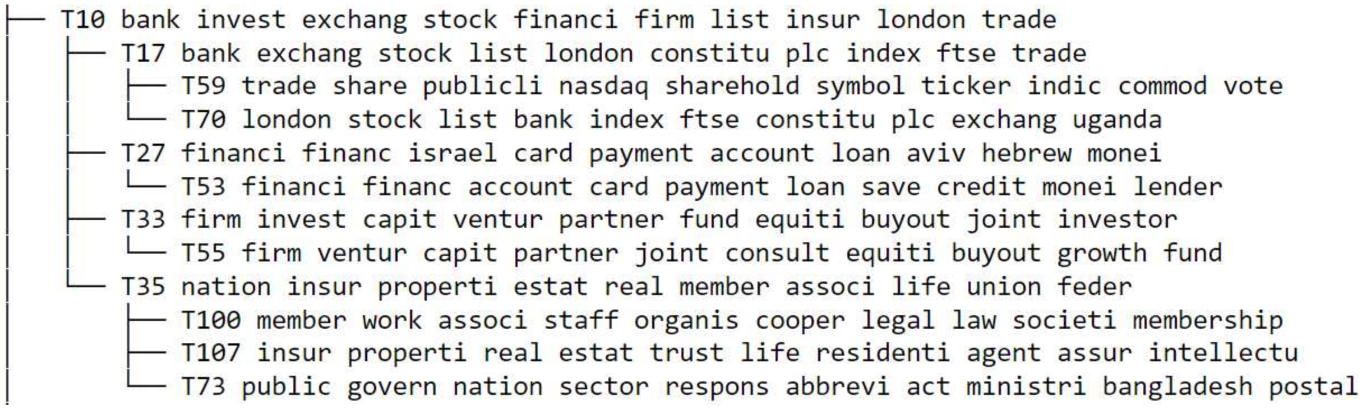
TABLE VI

AVERAGE SCORES OF TOPIC COHERENCE, TOPIC DIVERSITY, AND TOPIC QUALITY METRICS FOR HDDM ACROSS FOUR DATASETS AT TOP, MIDDLE, AND BOTTOM LEVELS.

	C_V	TD	TQ
level	mean (std)	mean (std)	mean (std)
top	0.50 (0.04)	0.96 (0.05)	0.48 (0.05)
middle	0.48 (0.03)	0.92 (0.10)	0.44 (0.05)
bottom	0.47 (0.04)	0.90 (0.12)	0.42 (0.06)

C. Hierarchical Structure Visualization

An advantage of unsupervised text modeling is the ability to visually present latent topics and examine the face validity of the outcomes. Such visualization serves as an exploratory tool for end-users to understand latent themes in the text corpus and facilitate effective content analysis. In Figure 3, we present the learned hierarchical topic structures from the Company Descriptions dataset. Due to space constraints, we display the hierarchy under the super-topic T10, focusing on financial markets indicated by top words such as ”bank,” ”invest,” ”exchange,” ”stock,” and ”financial.” Further, T10 is divided into four sub-branches: T17, T27, T33, and T35, each representing different topics. For instance, T17 may signify the stock market topic, with top words like ”exchange,” ”stock,” ”list,” ”index,” and ”trade,” while T27 could represent banking services, indicated by words such as ”card,” ”payment,” and



```
├── T10 bank invest exchang stock financi firm list insur london trade
│   ├── T17 bank exchang stock list london constitu plc index ftse trade
│   │   ├── T59 trade share publicli nasdaq sharehold symbol ticker indic commod vote
│   │   └── T70 london stock list bank index ftse constitu plc exchang uganda
│   ├── T27 financi financ israel card payment account loan aviv hebrew monei
│   │   └── T53 financi financ account card payment loan save credit monei lender
│   ├── T33 firm invest capit ventur partner fund equiti buyout joint investor
│   │   └── T55 firm ventur capit partner joint consult equiti buyout growth fund
│   └── T35 nation insur properti estat real member associ life union feder
│       ├── T100 member work associ staff organis cooper legal law societi membership
│       ├── T107 insur properti real estat trust life residenti agent assur intellectu
│       └── T73 public govern nation sector respons abbrevi act ministri bangladesh postal
│
```

Fig. 3. Topic hierarchy for super-topic T10. The top 10 words with the highest probability (after reranking) under each super-/sub-topic are shown. All words are shown in their stemmed forms. The T-number is the topic index; for example, T17 is the 17th topic in the model.

”loan.” In addition, T33 suggests the topic of corporate finance, with words like ”firm,” ”capital,” ”equity,” and ”buyout,” while T35 may pertain to life and real estate insurance, with words like ”nation,” ”insurance,” ”life,” and ”real estate.” Furthermore, these intermediate topics are further divided into sub-topics. For example, T17 (stock market) is subdivided into T59 (possibly representing the US market) and T70 (possibly representing the UK market).

In summary, in this section, we conducted experiments to evaluate the topic modeling capability of our proposed approach and demonstrated its efficacy relative to the state-of-the-art, in document modeling and topic quality tasks. To better understand the utility of the proposed hierarchical topic model, we next study its effectiveness in a downstream predictive modeling task in healthcare: patient diagnosis prediction.

VI. Application: Hierarchical Topic Modeling on Clinical Notes

Learning from complex, high-dimensional clinical data is a key challenge for healthcare enhancement. Uncovering insights from abundant clinical notes allows hospitals to develop better predictive capabilities, enabling them to identify high-risk patients and facilitate clinicians’ decision making [69], [70]. For example, predicting patients’ ICU length of stay based on early clinical notes helps hospitals to manage their resources [69]. Inferring patients’ diagnoses from early clinical notes can also help with interventions and healthcare policies.

In this section, we evaluate the practical utility of hierarchical topic modeling in extracting key information from clinical notes. In particular, we consider two clinical prediction tasks: **ICU length of stay prediction** and **patient diagnosis prediction**. Following prior literature [56], we use different types of clinical notes for different prediction tasks due to the nature of task. For the ICU length of stay prediction, we use ICU notes taken in the first 48 hours of a visit, including nurses’ notes and physicians’ notes, to predict the number of days remaining in the ICU. For the patient diagnosis prediction task, we use clinical notes taken in the first 24 hours of a visit, including ECG notes and radiology notes, of a patient’s admission to hospital to predict the patient’s diagnosis. Note that a patient may not necessarily be transferred to an ICU within the first 24 hours. Following the literature [56], we obtain 25 conditions commonly recorded in the MIMIC-III ICD-9 diagnosis table. Thus, the problem becomes a multi-label classification problem, wherein each patient is associated with multiple diagnoses.

A. Predictive Methods

Since topic modeling is an unsupervised text analysis approach, we first train the topic model on the training clinical notes, and then we obtain and infer the topic vectors of training and testing clinical notes. We then use a logistic regression model on the topic vectors as the supervised learning model. Therefore, the clinical task prediction becomes a two-stage learning process: first clinical notes are represented in low-dimensional topic distributions, and second, the learned topic features serve as input for a logistic regression model [19], [71]. In the experiment, we include one flat topic modeling approach (LDA) and one hierarchical topic modeling approach (hLDA) in our benchmarking.

In addition to the unsupervised topic modeling methods, we also consider two supervised deep learning approaches for a fair comparison. The first is a deep learning model Bi-LSTM as a baseline. In this model, the input words of a clinical note are represented as pre-trained GloVE word embeddings [72]. We keep all tokens in the vocabulary, removing only punctuation and numbers. We choose this model as a baseline because LSTM models have been used to analyze clinical notes in prior predictive analytics literature [73].

We also fine-tune a pretrained language model, BERT. BERT is a popular transformer-based architecture that has achieved state-of-the-art results across numerous natural language processing tasks [74], [75]. However, directly fine-tuning BERT with clinical notes is not possible, because BERT requires a fixed input sequence length (512 tokens), which is usually exceeded by clinical notes. To handle the maximum length problem, we adopt long-document strategy afforded by the RoBERTa model [76], which is a hierarchical transformer designed to handle long text. Using this model, the input text is divided into k segments. The segments are then fed into a frozen BERT layer to obtain the representations. The output of the BERT layer then serves as the input to a bidirectional

LSTM (Bi-LSTM) layer. As a result, the model is no longer constrained by the maximum document length.

Note that the purpose of this experiment is to examine the capability of our proposed approach, HDDM, in learning from complex, high-dimensional clinical data. The outcome of HDDM can be then integrated into sophisticated clinical prediction models such as Doctor AI [77] or models with multimodal data such as vital signs or lab tests [73], [78].

B. Evaluation Metrics

We use the area under the receiver operating characteristic (ROC) curve (AUC) for prediction evaluation. An ROC curve is a two-dimensional graph in which the true positive rate is plotted against the false positive rate for various decision threshold settings. The AUC score measures the area under the ROC curve, giving a single score for classifier model performance. The value of the AUC score ranges from 0.5 to 1; models with higher AUC scores yield better predictions. We split the dataset into 80% training, 10% validation and 10% testing data. We train and tune the models on the training and validation sets, respectively. We further evaluate the models on the testing set. We repeat the experiment 30 times by changing the data split random seed.

C. Task 1: ICU Length of Stay Prediction

This task aims to predict the number of days after a patient is transferred into the ICU. Length of stay has frequently been used as a measure of ICU resource use [79], and accurately predicting length of stay is important for ICU scheduling and resource management. It also aids decision making on treatments, interventions, and healthcare policies. Early clinical information, such as nursing notes, are particularly valuable, especially as information such as laboratory culture results may not be available at the beginning of a patient's ICU stay [79]. Thus, a dependable means of early ICU length stay prediction would have a major positive impact on ICU care.

Following [56], we frame the length of stay prediction as a classification problem with eight classes (six classes for each day after 48 hours after transferal to the ICU, one for stays of more than one week but fewer than two weeks, and one for stays over two weeks). We report the one-vs-rest multi-class classification performance in Table VII. HDDM consistently performs better than the baseline models. Overall, the macro-AUC score of HDDM is 0.743, while those of the state-of-the-art deep learning approaches, BERT and Bi-LSTM, are 0.679 and 0.646, respectively. Thus, HDDM makes a substantial and significant improvements. For example, HDDM improves over BERT predictions by 9.4% and Bi-LSTM by 15.0%. Similarly, HDDM outperforms its unsupervised counterparts LDA and hLDA by a large margin. We believe that there are two reasons why unsupervised HDDM with simple logistic regression outperforms state-of-the-art deep learning approaches. First, this unsupervised dimensionality-reduction approach is effective because it addresses the characteristics of clinical notes: noisy, unstructured, and high-dimensional. Second, clinical notes are usually very long, i.e., over 1,000 word tokens. Therefore, the BERT model is constrained by document length. Although we apply heuristics to address the long document concern, the BERT model is still prone to under/overfitting.

TABLE VII
AUC scores for ICU length of stay prediction performance, based on notes taken within the first 48 hours of transfer to the ICU. "Ratio" refers to the percentage of corresponding class labels. "Total" refers to the macro-AUC score.

# days	Ratio	LDA	hLDA	Bi-LSTM	BERT	HDDM
2-3	0.352	0.563	0.576	0.653	0.620	**0.660**
3-4	0.189	0.589	0.571	0.630	0.635	**0.664**
4-5	0.114	0.600	0.631	0.601	0.631	**0.711**
5-6	0.070	0.608	0.618	0.606	0.658	**0.682**
6-7	0.049	0.598	0.640	0.618	0.655	**0.767**
7-8	0.035	0.663	0.651	0.566	0.677	**0.768**
8-14	0.107	0.741	0.725	0.714	0.746	**0.806**
>14	0.083	0.726	0.733	0.778	0.810	**0.885**
Total		0.636	0.643	0.646	0.679	**0.743**

D. Task 2: Inferring Patient Diagnoses

We study another important clinical prediction task: patient diagnosis inference. This task may be crucial for healthcare outcomes. For example, inferring patients' diagnoses using early clinical notes such as radiology reports can help physicians to assess the severity of patients' illnesses and thus assist clinicians with decision making. In this experiment, we consider inferring patient diagnoses after 24 hours of hospitalization. We consolidate the dataset consisting of all clinical notes within 24 hours of admission to hospital (not necessarily to an ICU). Following the literature [56], we obtain 25 conditions commonly recorded in the MIMIC-III ICD-9 diagnosis table. Thus, as noted earlier, the problem becomes a multi-label classification problem, wherein each patient is associated with multiple diagnoses.

The patient diagnosis prediction results are presented in Table VIII, leading to the following findings. HDDM consistently outperforms the text analytics baselines both substantially and significantly. For example, across all prediction at admission tasks, HDDM outperforms BERT by 8.7% in terms of Macro-AUROC (0.796 vs. 0.732). HDDM also significantly outperforms LDA and hLDA (0.702 and 0.706, respectively). Moreover, we can see that the improvements of HDDM over benchmark methods are consistent across different diagnoses types. In particular, HDDM attains the best AUC on 24 out of 25 diagnosis class labels. Overall, the results suggest that HDDM is capable of accurately extracting clinically-relevant information from complex, high-dimensional clinical notes resulting in performance gains in topic modeling and downstream prediction tasks.

VII. Conclusion

Traditional topic modeling assumes that there are no topic hierarchies - however, many real-world textual contexts are comprised of concept hierarchies encompassing general and finer-grained subtopics. Failure to explicitly model hierarchical relationships may result in poor performance in topic analysis, and may also cascade into poorer performance in downstream predictive analytics tasks. Although Bayesian and

TABLE VIII
PATIENT DIAGNOSES USING NOTES TAKEN WITHIN 24 HOURS OF HOSPITALIZATION. PREVALENCE DENOTES PERCENTAGE OF COHORT WITH DIAGNOSIS.

Diagnosis	Prevalence	LDA	hLDA	LSTM	BERT	HDDM
Acute and unspecified renal failure	0.214	0.720	0.718	0.719	0.709	**0.799**
Acute cerebrovascular disease	0.073	0.850	0.852	0.889	**0.924**	0.915
Acute myocardial infarction	0.104	0.774	0.776	0.777	0.857	**0.887**
Cardiac dysrhythmias	0.321	0.722	0.722	0.684	0.754	**0.773**
Chronic kidney disease	0.134	0.712	0.721	0.719	0.729	**0.804**
Chronic obstructive pulmonary disease	0.130	0.663	0.675	0.638	0.640	**0.730**
Complication of surgical/medical care	0.208	0.650	0.659	0.672	0.679	**0.735**
Conduction disorder	0.072	0.718	0.722	0.693	0.660	**0.835**
Congestive heart failure	0.268	0.744	0.745	0.737	0.822	**0.840**
Coronary atherosclerosis and related	0.323	0.771	0.755	0.780	0.840	**0.863**
Diabetes mellitus with complications	0.095	0.726	0.729	0.701	0.711	**0.819**
Diabetes mellitus without complications	0.193	0.545	0.556	0.625	0.595	**0.635**
Disorder of lipid metabolism	0.290	0.621	0.624	0.711	0.683	**0.733**
Essential hypertension	0.419	0.594	0.591	0.630	0.585	**0.651**
Fluid and electrolyte disorder	0.269	0.635	0.633	0.703	0.666	**0.742**
Gastrointestinal hemorrhage	0.073	0.776	0.776	0.739	0.737	**0.865**
Hypertension with complications	0.132	0.711	0.718	0.691	0.699	**0.794**
Other liver disease	0.089	0.744	0.748	0.698	0.803	**0.845**
Other lower respiratory disease	0.052	0.585	0.586	0.646	0.636	**0.694**
Other upper respiratory disease	0.041	0.660	0.664	0.638	0.736	**0.826**
Pleurisy	0.087	0.636	0.649	0.662	0.663	**0.713**
Pneumonia	0.139	0.701	0.707	0.738	0.734	**0.807**
Respiratory failure	0.181	0.746	0.748	0.793	0.833	**0.865**
Septicemia	0.143	0.775	0.778	0.756	0.815	**0.863**
Shock	0.079	0.783	0.787	0.732	0.782	**0.872**
Macro-AUROC		0.702	0.706	0.711	0.732	**0.796**

deep learning-based hierarchical methods have been proposed, existing methods have limitations such as reliance on predefined hierarchies or lack of methods/functions for layer stacking. To address this research gap, we propose HDDM, a hierarchical topic model that organizes latent topics in a hierarchical structure using deep VAE and employs a new objective that can effectively learn super/sub-topic dependencies.

We conduct experimental analysis on several real-world datasets to evaluate the performance of HDDM. First, HDDM better fits the data, as demonstrated by lower perplexity scores across several datasets. Second, HDDM demonstrates its capability to learn distinguishable and diverse topics, as evidenced by metrics such as topic coherence and topic diversity. Third, topics learned via HDDM and used as input features for patient diagnosis prediction outperform other topic-modeling features and more advanced deep learning approaches.

As hierarchical topic modeling is particularly suitable for industry-level applications when data is plentiful (see Google's Rephil for example), this paper, which offers an effective and efficient solution to hierarchical neural topic modeling, has potential to make contributions to large web scale applications. Future work can also build upon the proposed framework and incorporate prior knowledge or document structural information to improve the document modeling capacity. Moreover, a growing body of research is focusing on identifying effective data analytics strategies for deriving actionable insights [20].

This research has limitations that can be improved in the future. First, HDDM infers the tree structure automatically from the text dataset. In some applications, researchers or practitioners may have prior knowledge about the topic hierarchies; future work can investigate incorporating domain knowledge to further enhance the quality of the learned topic hierarchies [80]. Second, in this work, we study four real-world datasets spanning consumer reviews, news articles, and clinical notes. Given the hierarchical nature of topics and documents, future work can empirically test HDDM in web-scale corpora.

REFERENCES

[1] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, Dec. 2014.

[2] L. Xu, C. Jiang, Y. Ren, and H.-H. Chen, "Microblog dimensionality Reduction—A deep learning approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1779–1789, Jul. 2016.

[3] Y. Zuo, C. Li, H. Lin, and J. Wu, "Topic modeling of short texts: A pseudo-document view with word embedding enhancement," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 972–985, Jan. 2023.

[4] W. Li, J. Yin, and H. Chen, "Supervised topic modeling using hierarchical dirichlet process-based inverse regression: Experiments on E-commerce applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1192–1205, Jun. 2018.

[5] Y. Yang and R. Subramanyam, "Extracting actionable insights from text data: A stable topic model approach." *MIS Quarterly*, vol. 47, no. 3, 2023.

[6] W. Ji, X. Meng, and Y. Zhang, "SPATM: A social period-aware topic model for personalized venue recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3997–4010, Aug. 2022.

[7] S. Mankad, S. Hu, and A. Gopal, "Single stage prediction with embedded topic modeling of online reviews for mobile app management," *The Annals of Applied Statistics*, vol. 12, no. 4, 2018.

[8] A. S. Abrahams, W. Fan, G. A. Wang, Z. Zhang, and J. Jiao, "An integrated text analytic framework for product defect discovery," *Production and Operations Management*, vol. 24, no. 6, pp. 975–990, 2015.

[9] Y. Zhang, P. Calyam, T. Joshi, S. Nair, and D. Xu, "Domain-specific topic model for knowledge discovery in computational and data-intensive scientific communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 1402–1420, Feb. 2023.

[10] Y. Miao, E. Grefenstette, and P. Blunsom, "Discovering discrete latent topics with neural variational inference," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2410–2419.

[11] A. Ahmed and E. P. Xing, "Seeking the truly correlated topic posterior-on tight approximate inference of logistic-normal admixture model," in *Artificial Intelligence and Statistics*. PMLR, 2007, pp. 19–26.

[12] D. Downey, C. Bhagavatula, and Y. Yang, "Efficient methods for inferring large sparse topic hierarchies," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 774–784.

[13] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations*, 2014.

[15] M. Kang, K. Ji, X. Leng, X. Xing, and H. Zou, "Synthetic aperture radar target recognition with feature fusion based on a stacked autoencoder," *Sensors*, vol. 17, no. 1, p. 192, 2017.

[16] P. Bachman, "An architecture for deep, hierarchical generative models," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[17] T. Zhao, R. Zhao, and M. Eskénazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,*, 2017, pp. 654–664.

[18] A. Abbasi, S. Sarker, and R. H. Chiang, "Big data research in information systems: Toward an inclusive research agenda," *Journal of the association for information systems*, vol. 17, no. 2, p. 3, 2016.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[20] Y. Yang, K. Zhang, and Y. Fan, "sdtm: A supervised bayesian deep topic model for text analytics," *Information Systems Research*, 2022.

[21] J. Chen, Y. Yang, and H. Liu, "Mining bilateral reviews for online transaction prediction: a relational topic modeling approach," *Information Systems Research*, vol. 32, no. 2, pp. 541–560, 2021.

[22] N. Zhong and D. A. Schweidel, "Capturing changes in social media content: a multiple latent changepoint topic model," *Marketing Science*, vol. 39, no. 4, pp. 827–846, 2020.

[23] J. Büschken and G. M. Allenby, "Sentence-based text analysis for customer reviews," *Marketing Science*, vol. 35, no. 6, pp. 953–975, 2016.

[24] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 577–584.

[25] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM (JACM)*, vol. 57, no. 2, pp. 1–30, 2010.

[26] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. JMLR Proceedings, vol. 15. JMLR.org, 2011, pp. 752–760.

[27] X. Wang and Y. Yang, "Neural topic model with attention for supervised learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1147–1156.

[28] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[29] Z. Chen, C. Ding, Y. Rao, H. Xie, X. Tao, G. Cheng, and F. L. Wang, "Hierarchical neural topic modeling with manifold regularization," *World Wide Web*, vol. 24, no. 6, pp. 2139–2160, 2021.

[30] M. Isonuma, J. Mori, D. Bollegala, and I. Sakata, "Tree-structured neural topic model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 800–806.

[31] Z. Chen, C. Ding, Z. Zhang, Y. Rao, and H. Xie, "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*, 2021, pp. 2343–2353.

[32] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[33] H. Zhang, B. Chen, D. Guo, and M. Zhou, "Whai: Weibull hybrid autoencoding inference for deep topic modeling," *arXiv preprint arXiv:1803.01328*, 2018.

[34] Z. Duan, D. Wang, B. Chen, C. Wang, W. Chen, Y. Li, J. Ren, and M. Zhou, "Sawtooth factorial topic embeddings guided gamma belief network," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2903–2913.

[35] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[36] Z. Zhang, M. Fang, L. Chen, and M. R. N. Rad, "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*, 2022, pp. 3886–3893.

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

[38] F. Bianchi, S. Terragni, and D. Hovy, "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)*, 2021, pp. 759–766.

[39] D. Wang, Y. Xu, M. Li, Z. Duan, C. Wang, B. Chen, M. Zhou *et al.*, "Knowledge-aware bayesian deep topic model," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 331–14 344, 2022.

[40] Z. Guo, Z. M. Zhang, S. Zhu, Y. Chi, and Y. Gong, "A two-level topic model towards knowledge discovery from citation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 780–794, Apr. 2014.

[41] S.-J. Shin and I.-C. Moon, "Guided HTM: Hierarchical topic model with dirichlet forest priors," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 330–343, Feb. 2017.

[42] Y. Meng, Y. Zhang, J. Huang, Y. Zhang, C. Zhang, and J. Han, "Hierarchical topic mining via joint spherical tree and text embedding," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 1908–1917.

[43] Y. Wang, J. Liu, Y. Huang, and X. Feng, "Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1919–1933, Jul. 2016.

[44] W. X. Zhao, J. Wang, Y. He, J.-Y. Nie, J.-R. Wen, and X. Li, "Incorporating social role theory into topic models for social media content analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 1032–1044, Apr. 2015.

[45] H. Azarbonyad, M. Dehghani, T. Kenter, M. Marx, J. Kamps, and M. de Rijke, "HiTR: Hierarchical topic model re-estimation for measuring topical diversity of documents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2124–2137, Nov. 2019.

[46] F. Ahmad, A. Abbasi, B. Kitchens, D. Adjeroh, and D. Zeng, "Deep learning for adverse event detection from web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2681–2695, 2020.

[47] J. Clark and F. Provost, "Matrix-factorization-based dimensionality reduction in the predictive modeling process: a design science perspective," 2016.

[48] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[49] Q. T. Tho, S. C. Hui, A. C. M. Fong, and T. H. Cao, "Automatic fuzzy ontology generation for semantic web," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 6, pp. 842–856, 2006.

[50] A. Kashyap, V. Hristidis, M. Petropoulos, and S. Tavoulari, "Effective navigation of query results based on concept hierarchies," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 4, pp. 540–553, 2010.

[51] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, "Topic and keyword re-ranking for lda-based topic modeling," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1757–1760.

[52] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," *arXiv preprint arXiv:1809.02687*, 2018.

[53] T. Nguyen and A. T. Luu, "Contrastive learning for neural topic model," *Advances in neural information processing systems*, vol. 34, pp. 11 974–11 986, 2021.

[54] N. Prakash, H. Wang, N. K. Hoang, M. S. Hee, and R. K.-W. Lee, "Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 621–631.

[55] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[56] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[57] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *arXiv preprint cs/0506075*, 2005.

[58] R. Qader, K. Jneid, F. Portet, and C. Labbé, "Generation of company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation," in *11th International Conference on Natural Language Generation*, 2018.

[59] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 977–984.

[60] A. Srivastava and C. Sutton, "Neural variational inference for topic models," in *Bayesian deep learning workshop, NIPS*, 2016.

[61] J. Boyd-Graber, D. Mimno, and D. Newman, "Care and feeding of topic models: Problems, diagnostics, and improvements," *Handbook of mixed membership models and their applications*, vol. 225255, 2014.

[62] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, May 2010, pp. 45–50.

[63] D. M. Blei and J. D. Lafferty, "Topic models," in *Text mining*. Chapman and Hall/CRC, 2009, pp. 101–124.

[64] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[65] A. Hoyle, P. Goel, A. Hian-Cheong, D. Peskov, J. Boyd-Graber, and P. Resnik, "Is automated topic model evaluation broken? the incoherence of coherence," *Advances in neural information processing systems*, vol. 34, pp. 2018–2033, 2021.

[66] A. M. Hoyle, R. Sarkar, P. Goel, and P. Resnik, "Are neural topic models broken?" in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 5321–5344.

[67] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

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

[69] H. Xu, W. Wu, S. Nemati, and H. Zha, "Patient flow prediction via discriminative learning of mutually-correcting processes," *IEEE transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 157–171, 2016.

[70] Y. Xu, H. Ying, S. Qian, F. Zhuang, X. Zhang, D. Wang, J. Wu, and H. Xiong, "Time-aware context-gated graph attention network for clinical risk prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[71] J. Zhu, A. Ahmed, and E. P. Xing, "Medlda: maximum margin supervised topic models," *the Journal of machine Learning research*, vol. 13, no. 1, pp. 2237–2278, 2012.

[72] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.

[73] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[76] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 838–844.

[77] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine learning for healthcare conference*. PMLR, 2016, pp. 301–318.

[78] Y. Li, P. Nair, X. H. Lu, Z. Wen, Y. Wang, A. A. K. Dehaghi, Y. Miao, W. Liu, T. Ordog, J. M. Biernacka *et al.*, "Inferring multimodal latent topics from electronic health records," *Nature communications*, vol. 11, no. 1, pp. 1–17, 2020.

[79] A. A. Kramer and J. E. Zimmerman, "A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay," *BMC medical informatics and decision making*, vol. 10, no. 1, pp. 1–16, 2010.

[80] Y. Yang, D. Downey, and J. Boyd-Graber, "Efficient methods for incorporating knowledge into topic models," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 308–317.

Yi Yang received the Ph.D. degree in Computer Science from Northwestern University in 2015, and bachelor's degree from University of Science and Technology of China in 2009. He is an Associate Professor in the Department of Information Systems and Operations Management (ISOM) at the Hong Kong University of Science and Technology (HKUST). His research interests relate to topic modeling, deep learning and large language models.

John P. Lalor received the Ph.D. degree from the University of Massachusetts Amherst in the College of Information and Computer Science. He is an Assistant Professor of IT, Analytics, and Operations at the University of Notre Dame. His research interests relate to NLP and text mining, including topic models and deep learning architectures for text sequence classification.

Ahmed Abbasi received his Ph.D. from the AI Lab at the University of Arizona. He is currently the Giovanini Endowed Chair at the University of Notre Dame. Ahmed serves as associate or senior editor at IEEE Intelligent Systems, ACM TMIS, and INFORMS ISR. He has received various grants and research awards for his work on machine learning and NLP including the IBM Faculty, IEEE Technical Achievement, and the INFORMS Design Science Awards. He is an IEEE senior member.

Daniel Dajun Zeng received a B.E. in operations research/economics with minor in CS from the University of Science and Technology of China, Hefei, and M.E. and Ph.D. degrees in industrial administration from Carnegie Mellon University. He is a professor at the Institute of Automation, Chinese Academy of Sciences. His research interests include machine learning, text mining, and social computing. He previously served as editor-in-chief at IEEE Intelligent Systems. Prof. Zeng is a fellow of IEEE.