Learning Representations from Medical Text for Effective Diagnoses and Knowledge Discovery

Anonymous ACL submission

Abstract

Discovering knowledge and effectively predicting target events are two main goals of medical data mining. However, few models can achieve them simultaneously. In this study, we investigated the possibility of discovering knowledge and predicting diagnosis at once via raw medical text. We proposed the Enhanced Neural Topic Model (ENTM), a variant of the neural topic model, to learn interpretable representations. We introduced the auxiliary loss set to improve the effectiveness of learned representations. Then, we used learned representations to train a softmax regression model to predict target events. As each element in representations learned by ENTM has an explicit semantic meaning, weights in softmax regression represent knowledge of whether an element is a significant factor in predicting diagnosis. We adopted two independent medical text datasets to evaluate our ENTM model. Results indicate that our model obtained better performance compared to the latest pretrained neural language models. Meanwhile, analysis of model parameters indicates our model can discover reliable knowledge from data.

1 Introduction

A large volume of patient information is recorded in text in the electronic health record (EHR) system. It is the primary evidence for doctors to realize disease characteristics and diagnoses. Extracting knowledge and diagnosing disease based on large, complex medical text data have been investigated for a long time in machine learning research (Chu et al., 2018; Koleck et al., 2019).

Intuitively, these two tasks are supposed to be tackled simultaneously. If a model achieves high disease predictive performance but is inexplainable, we will worry about whether the model makes decisions via sound inference evidence. If a model can extract knowledge from data but obtain poor disease predictive performance, we may suspect the reliability of the extracted knowledge. Doctors trust a model only when it can diagnose diseases accurately and it helps them understand the disease. However, recent studies usually tackled two tasks separately. Deep learning (DL) based studies typically obtained promising predictive performance but cannot provide any extra knowledge to doctors because of the black-box property (Payrovnaziri et al., 2020). On the contrary, knowledge discovery-oriented studies still use traditional models and ignore the DL techniques, which usually cannot obtain impressive disease predictive performance (Bellou et al., 2019; Shin et al., 2021).

The main goal of this study is to propose a method that can effectively predict patient diagnosis and discover knowledge at once via raw medical text. The term "effective" means our model needs to achieve comparable performance compared to the latest large-scale pretrained language models (PLM). In this study, we treat the predictive performance of large PLMs as the state-of-the-art (SOTA) performance of medical-text-based diagnosing tasks. The term "discover knowledge" indicates finding population-level characteristics that correlate to the occurrence of diseases. We regard discovered knowledge as reliable when in accord with current medical literature.

We proposed a two-phase model to achieve the goal of this study. The first phase is to learn representations of medical text. The second phase adopts the softmax regression method to predict diagnoses via representations. Of note, we adopted the assumption that linear models are interpretable. Softmax regression is one of the most widely used linear models to make predictions and discover knowledge in medical research. If the weight of a feature learned by the model is statistically significantly larger than zero, the feature may be a risk factor for the disease. Otherwise, it may be a protective factor.

Learning effective text representations whose el-

ement has explicit semantic meanings is the core challenge in achieving our goal. Only when each element in the representation has explicit meaning analyzing the feature weight learned by softmax regression is meaningful and discovering knowledge is possible. Topic model, e.g., Latent Dirichlet Allocation (LDA), can compress an unstructured text into an interpretable representation, a.k.a., the document-topic distribution (Blei et al., 2001). Each element in the representation can be mapped to a topic-word distribution, which can be endowed with an explanation by summarizing the word frequency. The element value indicates the importance of the corresponding topic-word distribution in a document. However, the representation is usually not effective enough because inferencing the true posterior of a complex topic model is intractable due to the high dimension integrals (Blei et al., 2017; Miao et al., 2017).

We introduced the enhanced neural topic model (ENTM) to extract more effective representations than LDA without loss of interpretability. The main principle of the ENTM is to use neural variational inference to model the posterior probability of an LDA. Besides, we introduced three auxiliary losses to leverage contrastive learning, knowledge distillation, and topic diversity information to train the model more effectively. Previous studies have demonstrated that the LDA can obtain unbiased and low variance representations using the neural variational inference reparameterization method (Miao et al., 2017; Nguyen and Luu, 2021). However, these models did not detailly investigate their predictive potential. To our knowledge, these models were not applied in medical research, either.

We conducted experiments to evaluate representations learned by the ENTM via disease diagnosing tasks on two medical datasets. Experimental results indicated that the representation learned by the ENTM is as effective as the representations learned by the latest PLMs as they achieved comparable performance. Meanwhile, disease knowledge discovered from topic-word distributions and softmax weights conformed to clinical literature and were advocated by clinical professionals.

2 Related Work

The emergence of DL technology inspired studies to propose neural network (NN) based interpretable models. They adopt NN to generate the weight of each feature to interpret predictions, a.k.a., the attention mechanism. The main difference between these prediction models and traditional linear models is that the weight of a feature is sample-specific in NN models, while the weight of a feature is a scalar shared in different samples in traditional linear models. RETAIN is the first model to generate feature weights to interpret its prediction (Choi et al., 2016). Its design was widely inherited in subsequent studies (Yin et al., 2019; Zhang et al., 2019; Ye et al., 2021; Zhang et al., 2020). Although useful, these studies only accept structured data to conduct interpretable predictions. Of note, structuring medical text inevitably introduces errors and biases. Sometimes it is even infeasible because of patient heterogeneity or lack of tools. Meanwhile, these studies did not systematically discuss or investigate the soundness of model-generated feature weights. Recent studies demonstrated that NN-based feature weights are unstable. Feature weights only have weak relations to other feature importance measures and cannot be treated as reliable explanations, or knowledge (Jain and Wallace, 2019; Kim et al., 2020).

The development of neural generative models provides new potential to topic models. Kingma *et al.* demonstrated it is possible to use a parameterized neural variational inference framework to simulate the posterior whose evidence is intractable (Kingma et al., 2014). Miao *et al.* proposed the neural variational document model (NVDM) and its variants, which can be regarded as a NN parameterized version of LDA, to generate the representations of documents (Miao et al., 2017). Srivastava *et al.* proposed the prodLDA independent to the NVDM (Srivastava and Sutton, 2017). They constructed a Laplace approximation to the Dirichlet prior, making their prodLDA model interpretable, while the NVDM is an uninterpretable model. Dieng *et al.* introduced logistic-normal distribution to generate document-topic distribution, and they applied word embeddings to optimize the performance of the model (Dieng et al., 2020).

3 Methodology

Fig. 1 describes our ENTM model, which can be regarded as a revision of the NVDM (Miao et al., 2016). We revised the word sampling strategy in this study to make our model explainable, while the original NVDM is an inexplainable model. Meanwhile, we introduced three auxiliary losses to train the model more effectively.

3.1 Neural Topic Model

Neural Topic Model (NTM) is the core of the ENTM. The NTM generally followed the framework of LDA in modeling text (Blei et al., 2001). Specifically, it presumes each word w_{dn} in a document d from corpus D is generated independently via the following process:

$$z_{dn} \sim \text{Categorical}(h_d) \tag{1}$$

$$w_{dn} \sim \text{Categorical}(\theta_{z_{dn}}) \tag{2}$$

where $d \in D$ is the document index, $n \in N_d$ is the word index, N_d indicates the word number in document d, and x_d denotes bag-of-words representation of a document. $h_d \in \mathbb{R}^T$ indicates the document-topic distribution over document d, z_{dn} indicates the topic of w_{dn}, $\theta_{z_{dn}} \in \mathbb{R}^V$ indicates the topic-word distribution over the topic z_{dn}, T and V indicate the topic number and vocabulary size, respectively. $\text{Categorical}(\cdot)$ is a categorical distribution. We treat θ as parameters and h as hidden variables. The log-likelihood of D is:

$$\log p(D;\theta) = \log \prod_{d=1}^{D} \int p(h) \prod_{n=1}^{N_d} (h\theta)_{w_{dn}} dh \tag{3}$$

It is challenging to optimize Eq. 3 as its integral is intractable. We sidestep the integral with variational inference. Specifically, we introduce variational distributions $q(h|\, d;\varphi)$ that approximate the posterior $p(h|\, d;\theta)$, where φ are the variational parameters. Equipped with the neural variational inference framework, the log-likelihood can be reorganized as (Blei et al., 2017):

$$\begin{aligned}\log p(D;\theta) = \text{KL}(q(h|\, d;\varphi)||p(h|\, D;\theta)) \\ + \text{ELBO}(q|\, d;\theta,\varphi)\end{aligned} \tag{4}$$

$$\begin{aligned}\text{ELBO}(q|\, d;\theta,\varphi) = \mathbb{E}_q(\log p(d|\, h;\theta)) \\ - \text{KL}(q(h|\, d;\varphi)||p(h))\end{aligned} \tag{5}$$

where KL indicates the Kullback-Leibler divergence, the ELBO (evidence lower bound) can be regarded as a lower bound of log probabilities. In this study, we optimize Eq.4 indirectly by optimizing ELBO (and omit the KL divergence term in Eq.4). It usually works well in practice, although theoretical correctness has not been justified yet.

Note $\log p(d|\, h;\theta)$ is irrelevant to $q(h|d\,;\varphi)$, the first term in Eq. 5 can be rewritten as:

$$\mathbb{E}_q(\log p(d|\, h;\theta)) = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \log(h_d\theta)_{w_{dn}} \tag{6}$$

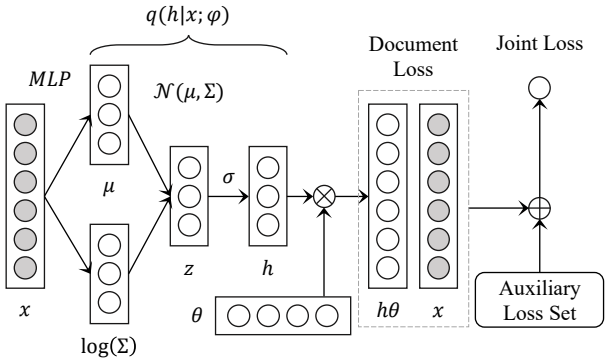

Figure 1: Model Diagram

We presume the prior $p(h)$ is a standard Gaussian distribution with a softmax function. Previous studies also used the log-normalize function or Laplace approximation to generate normalized topic distribution (Dieng et al., 2020). We followed the design of (Kingma et al., 2014) to introduce an inference network. $q(h|\, d;\varphi)$ is modeled as a Gaussian distribution with a softmax function as well. μ, $\Sigma_d = f_\varphi(x_d)$ are location and covariance matrix of $q(h|\, d;\varphi)$, where $f_\varphi(\cdot)$ is a NN parameterized by φ. To evaluate the KL divergence analytically, we use the KL divergence of unnormalized $p(h)$ and $q(h|\, d;\varphi)$ distributions as the proxy of the true KL divergence. The proxy KL divergence follows:

$$\begin{aligned}\text{KL}(\mathcal{N}(\mu,\, \Sigma_d)||\mathcal{N}(0,\mathbf{I})) = \\ -\frac{1}{2}(T - \mu\mu^T - tr(\Sigma_d) + \log|\Sigma_d|)\end{aligned} \tag{7}$$

The document loss function of a minibatch $\mathcal{B}$ can be summarized as:

$$\begin{aligned}\mathcal{L}_d = -\sum_{d=1}^{\mathcal{B}}\sum_{n=1}^{N_d} \log(h_d\theta)_{w_{dn}} - \\ \sum_{d=1}^{\mathcal{B}} \frac{1}{2}(T - \mu\mu^T - tr(\Sigma_d) + \log|\Sigma_d|)\end{aligned} \tag{8}$$

We can optimize document loss via a gradient-based method. The training process is summarized in Algorithm 1.

3.2 Auxiliary Loss Set

As Eq.8 is just an approximation of the true log-likelihood, and the training dataset may not be sufficiently large, we argue that the learned parameters may not be effective and robust. We introduced three auxiliary losses to improve the model.

Algorithm 1 NTM Training Process

Input: Corpus D
Output: φ, θ

Random initialize : φ, $\beta \in \mathbb{R}^{T \times V}$
while φ and β not converge **do**
  for iteration $d = 0$ to D **do**
    θ_d = softmax(β_d)
  end for
  Sample minibatch $\mathcal{B}$ from D
  for each x_d in $\mathcal{B}$ **do**
    Compute μ, $\Sigma_d = f_\varphi(x_d)$
    Sample $h_d = \mathrm{softmax}(\mathcal{N}(\mu, \Sigma_d))$
  end for
  for each word w_{dn} in document d **do**
    Compute $p(w_{dn}) = (h_d\theta)_{w_{dn}}$
  end for
  Compute gradient of $\mathcal{L}_d$ according to Eq.8
  Update parameters β, φ
end while
return φ, θ

3.2.1 Contrastive Loss

The document label, e.g., diagnosis, is available in the medical document but is not used. The label may improve the NTM by providing additional high-level similarity information. In this study, we introduce a contrastive learning-based loss term to utilize the information (Chen et al., 2020). We encourage the representations with the same label to become similar and those with different labels to become dissimilar.

$$\mathcal{L}_c = \sum_{d_1}^{\mathcal{B}} \sum_{d_2 \neq d_1}^{\mathcal{B}} (\mathbb{I}(y_{d_1} \neq y_{d_2}) - \mathbb{I}(y_{d_1} = y_{d_2})) \left(\frac{h_{d_1} h_{d_2}^T}{\|h_{d_1}\| \, \|h_{d_2}\|}\right) \tag{9}$$

Where d_1, d_2 are document indices. y_{d_1}, y_{d_2} are document labels. $\mathbb{I}(\cdot)$ is indicator function.

3.2.2 Knowledge Distillation Loss

We developed a simple method to transfer the knowledge of PLM into our ENTM model (Hinton et al., 2015). Specifically, given arbitrary document pairs, we argue that the representation similarity learned by ENTM should be the same as the representation similarity generated by PLM. According to the assumption, we introduce the knowledge distillation loss (Eq. 10).

$$\mathcal{L}_k = \sum_{d_1}^{\mathcal{B}} \sum_{d_2 \neq d_1}^{\mathcal{B}} \left[\frac{r_{d_1} r_{d_2}^T}{\|r_{d_1}\| \, \|r_{d_2}\|} - \frac{h_{d_1} h_{d_2}^T}{\|h_{d_1}\| \, \|h_{d_2}\|} \right]^2 \tag{10}$$

where r_{d_1}, r_{d_2} are document representations generated by a PLM.

3.2.3 Topic Diversity Loss

Existing studies have found that optimizing log-likelihood negatively correlates with the quality of extracted topic (Chang et al., 2009). To extract meaningful semantic topics, we introduce a topic diversity loss that encourages different topics to become dissimilar (Eq. 11).

$$\mathcal{L}_t = \sum_{t_1}^{T} \sum_{t_2 \neq t_1}^{T} \left(\frac{\theta_{t_1} \theta_{t_2}^T}{\|\theta_{t_1}\| \, \|\theta_{t_2}\|}\right) \tag{11}$$

3.2.4 Joint Loss

In summary, the joint loss follows:

$$\mathcal{L} = \alpha\mathcal{L}_d + \beta\mathcal{L}_c + \gamma\mathcal{L}_k + \delta\mathcal{L}_t \tag{12}$$

where α, β, γ, δ are loss weights. We call the NTM equipped with three auxiliary losses the ENTM. The sum of the four hyperparameters should equal one, i.e., $\alpha + \beta + \gamma + \delta = 1$

4 Experiment

4.1 Experiment Settings

4.1.1 Dataset and Preprocessing

In this study, we experimented on two datasets. The first dataset is a Chinese EHR dataset from a tertiary psychiatric hospital in China (HZSPH), which includes clinical narratives of 1,500 patients with anxiety, depression, or bipolar affective disorder admitted to the hospital between 2019 and 2021. The second dataset is the MIMIC-III, a public dataset that includes complete EHR data of 58,997 hospitalizations of patients who visited the intensive care unit of Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). We described sample data in the Supplementary Material.

As one goal of this study is to predict the diagnosis of patients, we only utilize the information collected in the admission phase. Other information, e.g., diagnosis information and summary of hospital course, was removed to avoid leaking the label. We adopted the pkuseg toolkit to tokenize the

	HZSPH	MIMIC-III
Samples	1,463	8,827
Avg. # tokens	2,245	1,521
Max. # tokens	3,563	5,467
Min. # tokens	374	320
Class number	3	10

Table 1: Data Statistics

HZSPH dataset and the tokenizer of Deberta to tokenize the MIMIC-III dataset (Luo et al., 2019; He et al., 2021). We used the TF-IDF method to select the most important 10,000 tokens within the two datasets to construct bag-of-words representations. We adopted MacBert and Longformer, which are SOTA PLMs in Chinese and English, to generate the representations of documents in HZSPH and MIMIC-III datasets, respectively (Cui et al., 2020; Beltagy et al., 2020). Documents that are longer than the maximum length of a PLM were truncated from the beginning because the information was written in the order of importance.

Patients in the MIMIC-III dataset typically have multiple diagnoses, and the distribution of diagnoses is highly imbalanced. We only reserve patients whose primary diagnosis is in the top ten frequent diagnoses. We only use the primary diagnosis as the label to formulate our task as a multi-class prediction task. We did not filter the HZSPH dataset because it is balanced. The statistics of two preprocessed datasets are in Table 1.

4.1.2 Implementation

The model training process can be divided into two phases. In the first phase, we trained the ENTM according to the method we introduced in the last section. In the second phase, we locked the parameter in the ENTM and utilized the representations generated by the ENTM to predict the corresponding diagnosis via a softmax regression classifier.

We investigated the interpretability and the potential of discovering knowledge of ENTM by analyzing topic-word distribution parameters and feature weights of the softmax regression. Considering the space limit, we only described topic-word distributions and feature weights of HZSPH datasets in this paper.

We used the grid search strategy to find the best loss weights. We used the five-fold cross-validation approach to fully utilize the data and reported the mean performance of the cross-validation experiments to evaluate the model. We set the topic number to ten according to the perplexity analysis. The source code was released in the Supplementary Material.

4.1.3 Baseline Models

(1) LDA (Blei et al., 2001). LDA is a topic model that can extract interpretable document representations. (2) sLDA (Mcauliffe and Blei, 2007). Supervised LDA (sLDA) is an LDA variant that utilizes document labels. (3) NVDM (Miao et al., 2016). NVDM is an LDA variant that utilizes a neural variational inference framework. (4) MacBert (Cui et al., 2020). MacBert is a variant of BERT that introduces a novel masking strategy. (5) Deberta (He et al., 2021). Deberta is a variant of BERT that utilizes the disentangled attention mechanism. (6) Longformer (Beltagy et al., 2020). Longformer is a variant of BERT that accepts a maximum of 4,096 tokens, while most BERT models only accept a maximum of 512 tokens. Deberta, Longformer is only available in English, and MacBert is only available in Chinese. All PLMs were fine-tuned via our datasets.

4.1.4 Metrics

We used accuracy and (macro-) F1 to evaluate the predictive performance. Perplexity was used to determine the number of topics (Blei et al., 2001). We additionally used topic coherence to evaluate topic quality quantitatively. We chose the normalized pointwise mutual information (NPMI) to evaluate topic coherence as it was widely used in previous studies (Eq. 13) (Aletras and Stevenson, 2013).

$$\text{NPMI} = \frac{2}{TN(N-1)} \sum_{t=1}^{T} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\log \frac{p(w_{ti}, w_{tj})}{p(w_{ti})p(w_{tj})}}{-\log p(w_{ti}, w_{tj})} \quad (13)$$

where w_t means the Top N tokens in topic t. We set N to ten. $p(w_{ti})$ indicates the proportion that w_{ti} is in a document, $p(w_{ti}, w_{tj})$ indicates the proportion that both w_{ti} and w_{tj} are in a document.

4.2 Predictive Performance

Table 2 shows the predictive ability of the ENTM and baselines. Models' performances on HZSPH are systematically better because the MIMIC-III has ten classes to predict, while the HZSPH dataset only has three classes. It is not surprising that the performances of LDA and sLDA are worse

	HZSPH		MIMIC-III	
	Acc.	F1	Acc.	F1
LDA	0.71	0.70	0.54	0.27
sLDA	0.73	0.71	0.55	0.28
MacBert	0.85	**0.86**	-	-
Deberta	-	-	0.67	0.48
Longformer	-	-	0.74	0.56
NVDM	0.75	0.73	0.56	0.30
NTM	0.76	0.75	0.57	0.30
NTM+CL	0.83	0.84	0.72	0.56
NTM+KL	0.79	0.79	0.60	0.44
NTM+TL	0.78	0.78	0.59	0.42
ENTM-512	0.84	0.84	0.68	0.49
ENTM	**0.86**	**0.86**	**0.75**	**0.59**

Table 2: Prediction Performance

than other models, as inferencing posterior is difficult. PLM-based models obtained significantly better performance than topic models. The ENTM obtained better performance than PLMs in both datasets, demonstrating our ENTM is efficient in diagnosing via raw medical text.

We noticed that the MacBert achieved comparable, while the Deberta obtained significantly worse performance than the ENTM. This phenomenon may be attributed to the data characteristics. We argue the vital information of HZSPH is recorded at the beginning of the text, while the vital information of MIMIC-III is relatively distributed evenly. Therefore, the data truncation procedure significantly deteriorates the performance of Deberta.

4.3 Ablation Study

We investigated the effectiveness of our design by conducting ablation studies (Table 2). The NVDM and NTM obtained comparable performance in both datasets. The NTMs trained with different losses, i.e., contrastive loss (CL), knowledge distillation loss (KL), and topic diversity loss (TL), obtained performance improvement independently, while improvements in introducing KL and TL are relatively marginal, and the improvement of introducing CL is significant. The ENTM obtained the best performance, indicating the combination of auxiliary losses can further improve representations' effectiveness. The ENTM-512, i.e., a variant of ENTM that only uses the first 512 tokens of a document, obtained comparable performance to Deberta and MacBert, indicating representations generated by our ENTM are as effective as those learned by the latest PLMs in the same experimental setting.

	HZSPH	MIMIC-III
LDA	N/A	N/A
sLDA	N/A	N/A
NTM	0.104	0.093
NTM+CL	0.181	0.182
NTM+KL	0.131	0.112
NTM+TL	0.135	0.109
ENTM	**0.197**	**0.192**

Table 3: Topic Coherence

4.4 Topic Coherence

Table 3 described the topic coherence of our model and baselines. We failed to calculate the NPMI of LDA and sLDA because a part of Top N token pairs w_{ti} and w_{tj} never co-exist in a document, which causes numerical error in calculating NPMI. These results indicated that the quality of some topics extracted by LDA and sLDA was poor.

All NTM-based models obtained NPMI scores successfully. The incorporation of CL significantly improved the quality of topics. The NTM+KL and NTM+TL models also obtained higher NPMI than the NTM model, though the improvement is marginal. The ENTM obtained the best NPMI in two datasets, indicating that all three auxiliary losses are effective in learning better topic-word distributions, while the combination of losses can further improve the quality of topics.

4.5 Knowledge Discovery Ability

Fig. 2 describes the result of the weight matrix of the softmax regression and the topic-word distribution of an HZSPH experiment. The larger the weight, the stronger the positive/negative influence of a topic in diagnosing a disease. We select 15 tokens with a high occurrence probability for each topic to explain their semantic meanings. We only investigated topics one, two, five, seven, and nine to reveal the characteristics of bipolar disorder, depression, and anxiety because they are the strongest factors.

The five topic-word distributions have significant differences. Each high-frequency token list can be regarded as a patient group depicting typical characteristics of the three diseases. We can easily find that topics one, two, and nine reflect a patient group that feels unsafe. They also described three

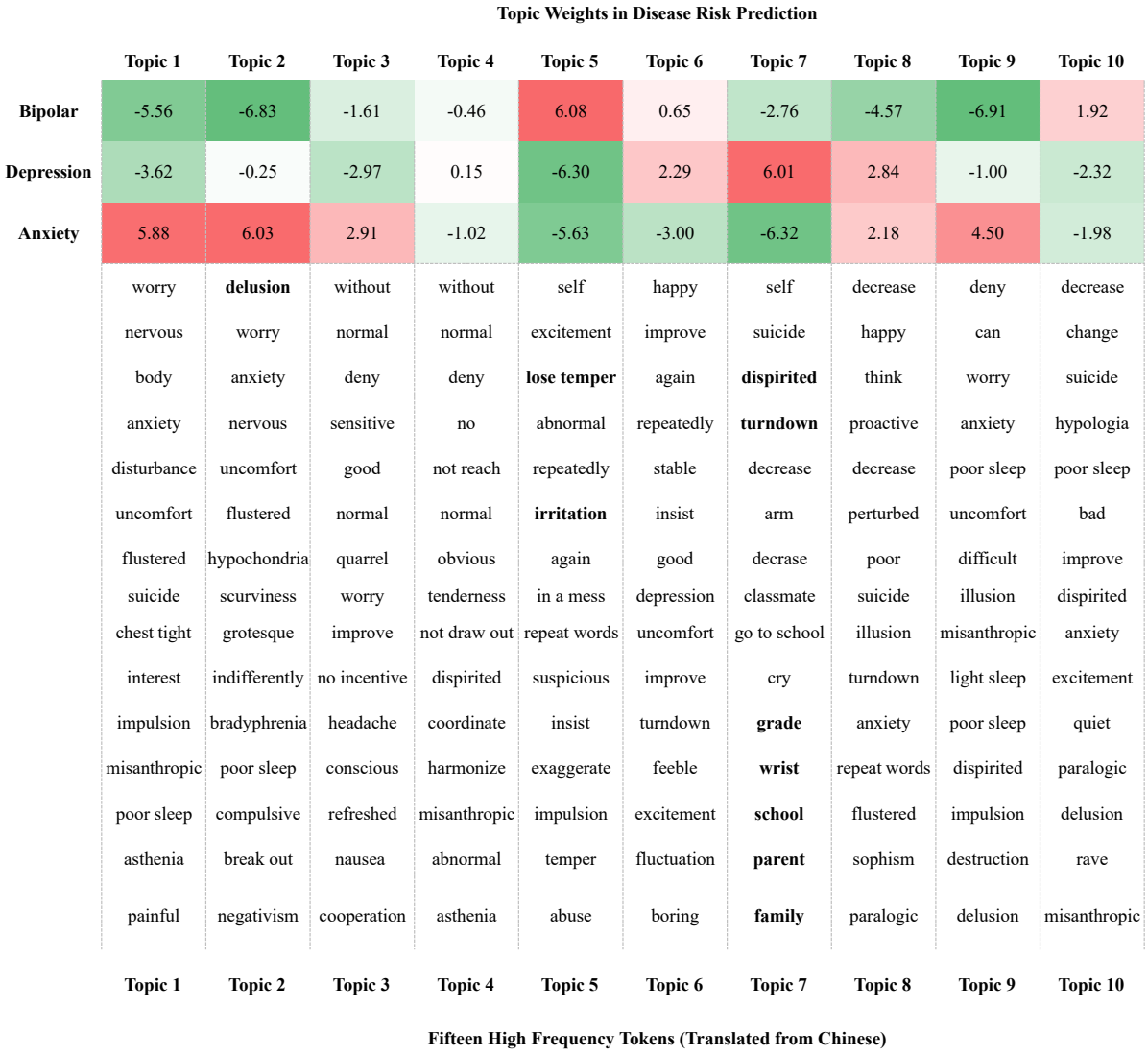

Figure 2: Knowledge Discovery Analysis

subtypes of anxiety as token lists are different. For example, the word "delusion" was first placed in topic two but was not included in topic one. Topic five revealed the symptom of mania as we can find a word such as "lose temper" and "irritation" in the list. Topic seven indicated a low-mood patient group as we found "dispirited" and "turndown" in the list. Another characteristic of the token list of topic seven is that it included tokens related to body parts, e.g., "wrist", and tokens related to students, e.g., "classmates" and "parents". These tokens are not included in other lists. We found these tokens were included because many depression patients are students and typically had self-harm records. Although the patient pattern characteristic needs to be interpreted carefully, it reveals the possibility that teenagers may be the most vulnerable population group to depression in China. All the characteristics discovered in this subsection accord with current clinical literature and are advocated by our clinical collaborators (Association, 2013). Therefore, we argue that the found characteristics can be regarded as reliable knowledge.

4.6 Influence of Loss Weights

Fig. 3 describes the performance of ENTM with different loss weights combinations. We set the weight of β, γ, δ to 0, 0.05, 0.1, and 0.15, and implemented 64 five-fold cross-validation experiments in two datasets, respectively. Each point in the figure reports the average accuracy difference between the ENTM with the corresponding loss weight set and the original NTM, i.e., $\text{acc}_{ENTM} - \text{acc}_{NTM}$. The more yellow the point, the more the improvement. The influence of three auxiliary losses is relatively insensitive to weight

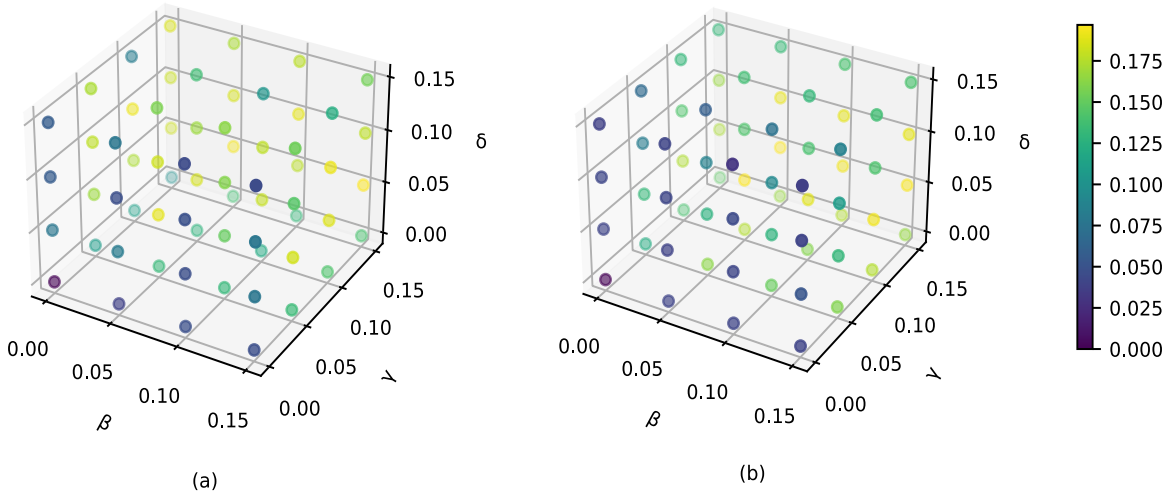

Figure 3: Influence of loss weight. (a) Accuracy difference on HZSPH dataset. (b) Accuracy difference on MIMIC-III dataset.

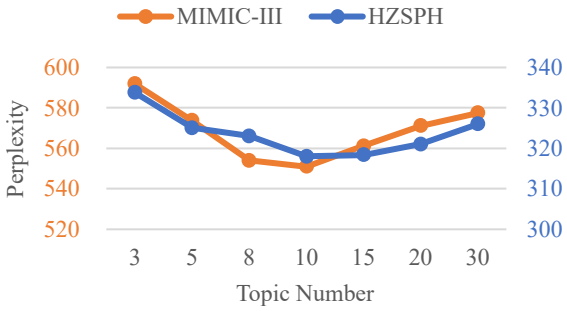

Figure 4: Perplexity

values as the ENTM model was improved in all combinations. The CL loss plays the main role in improving model performance. However, the other two losses are not negligible as the incorporation of KL and TL improved our model further. These experimental results indicated that the improvement effect of three auxiliary losses is obvious and stable.

4.7 Perplexity

Fig. 4 shows the perplexity of ENTM with different topic numbers. In the HZSPH dataset, perplexities decrease with the increase of topic numbers when they are less than ten. Perplexities gradually increase with the increase of topic numbers when they are larger than 15. Perplexities are basically unchanged when the topic number is between ten to 15. In the MIMIC-III dataset, perplexities decrease with the increase of topic numbers when they are less than ten. Perplexities gradually increase with the increase of topic numbers when they are larger than ten. Therefore, we chose ten as the topic number to conduct experiments.

5 Conclusion

Our ENTM inherits the advantage of the LDA in knowledge discovery ability. We introduced the neural variational inference framework and auxiliary loss set to improve the representation generated by ENTM. The predictive performance of the ENTM was evaluated on two independent EHR datasets. The experiment result indicates that ENTM obtained better performance than the latest PLMs with significantly fewer computational resources, and our model design is effective. The knowledge discovery analysis demonstrates that the ENTM can extract representative patient characteristics from raw data, and the characteristics accord with current medical literature, which is beyond the ability of PLMs. In summary, we achieved the goal of predicting patient diagnosis effectively and discovering knowledge at once via raw medical.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Int. Conf. Comput. Semantics*, pages 13–22.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.).*

Vanesa Bellou, Lazaros Belbasis, Athanasios K Konstantinidis, Ioanna Tzoulaki, and Evangelos Evangelou. 2019. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ*, 367.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Adv. Neural. Inf. Process. Syst.*, volume 14.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Adv. Neural. Inf. Process. Syst.*

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learning*.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Adv. Neural. Inf. Process. Syst.*, volume 29.

Jiebin Chu, Wei Dong, Kunlun He, Huilong Duan, and Zhengxing Huang. 2018. Using neural attention networks to detect adverse medical events from electronic health records. *J. Biomed. Inf.*, 87:118–130.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings Assoc. Comput. Linguistics EMNLP*, pages 657–668.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.*, 8:439–453.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Int. Conf. Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Conf. North Am. Chapter Assoc. Comput. Linguist.*, pages 3543–3556.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Sci. Data*, 3(160035).

Junetae Kim, Sangwon Lee, Eugene Hwang, Kwang Sun Ryu, Hanseok Jeong, Jae Wook Lee, Yul Hwangbo, Kui Son Choi, Hyo Soung Cha, et al. 2020. Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the korean diabetic disease setting. *J. Med. Internet Res.*, 22(12):e18418.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Adv. Neural. Inf. Process. Syst.*, volume 27.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J. Am. Med. Inform. Assoc.*, 26(4):364–379.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *arXiv preprint arXiv:1906.11455*.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. In *Adv. Neural. Inf. Process. Syst.*, volume 20.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Int. Conf. Mach. Learning*, volume 70, pages 2410–2419.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. *Neural Variational Inference for Text Processing*. In *Int. Conf. Mach. Learning*, pages 1727–1736.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Adv. Neural. Inf. Process. Syst.*, volume 34, pages 11974–11986.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J. Am. Med. Inform. Assoc.*, 27(7):1173–1185.

Sheojung Shin, Peter C Austin, Heather J Ross, Husam Abdel-Qadir, Cassandra Freitas, George Tomlinson, Davide Chicco, Meera Mahendiran, Patrick R Lawler, Filio Billia, et al. 2021. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure*, 8(1):106–115.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Int. Conf. Learning Representations*.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *ACM SIGKDD Int. Conf. Inf. Knowledge Management*, pages 2414–2423.

Jiayu Yin, Xiang Lu, Zhiyuan Qian, Weiting Xu, and Xiang Zhou. 2019. New insights into the pathogenesis and treatment of sarcopenia in chronic heart failure. *Theranostics*, 9(14):4019–4029.

Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson. 2020. Inprem: An interpretable and trustworthy predictive model for healthcare. In *ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, pages 450–460.

Xianli Zhang, Buyue Qian, Yang Li, Changchang Yin, Xudong Wang, and Qinghua Zheng. 2019. Knowrisk: An interpretable knowledge-guided model for disease risk prediction. In *IEEE Int. Conf. Data Mining*, pages 1492–1497.