Towards Improving Topic Models with the BERT-based Neural Topic Encoder

Anonymous ACL submission

Abstract

Neural Topic Models (NTMs) have been popular for mining a set of topics from a collection of corpora. Recently, there is an emerging direction of combining NTMs with pre-trained language models such as BERT, which aims to use the contextual information of BERT to help train better NTMs. However, existing works in this direction either use the contextual information of pre-trained language models as the input of NTMs or align the outputs of the two kinds of models. In this paper, we study how to build deeper interactions between NTMs and pre-trained language and propose a BERT-based neural topic encoder, which deeply integrates with the transformer layers of BERT. Our proposed model encodes both the BoW data and the sequence of words of a document, which can be complementary to each other for learning a better topic distribution for the document. The proposed encoder is a better alternative to the ones used in existing NTMs. Thanks to the in-depth integration with BERT, extensive experiments show that the proposed model achieves the state-of-art performances the comparisons with many advanced models.

1 Introduction

As an unsupervised text analysis technique, topic modeling has been used in various Natural Language Processing (NLP) tasks (Boyd-Graber et al., 2017; Wang et al., 2019). Conventional Bayesian topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are usually based on Gibbs and variational inference, which show signs of fatigue in the face of big data and deep learning (Huynh et al., 2020; Zhao et al., 2021a).

Recent empirical studies show neural topic models (NTMs) can both successfully discover high-quality topics and document representations (Miao et al., 2017; Srivastava and Sutton, 2017; Zhao et al., 2021b; Duan et al., 2021). Most of NTMs can be viewed as the extensions of LDA on the framework of variational autoencoders (VAEs) where a latent topic distribution inference network (encoder) and a document generative model (decoder) are employed. The encoder takes the Bag-of-Word (BoW) vector of a document as input and infers the posterior distribution of the latent topics and the decoder takes the samples from the posterior distribution as input to reconstruct the BoW representation. Generally speaking, NTMs are trained by maximizing its evidence lower bound (ELBO) using gradient ascent, resulting in better flexibility and scalability than conventional topic models.

Similar to conventional TMs, NTMs typically take the sparse BoW vector as input, which disregards the syntactic and semantic relationships among the words in a document, leading to downgraded performance. To solve the problem, it is reasonable to train NTMs by leveraging the knowledge from pre-trained language models. Recently, pre-trained language models such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) have been successful in various NLP tasks. The core idea behind such popularity is the predominant pre-train and fine-tune paradigm. Pre-trained on huge text, such models can serve as a powerful encoder that outputs contextual semantic token embeddings. Fine-tuned according to the new task, these models have advanced state-of-the-art performance across many tasks (Liu et al., 2019; Rogers et al., 2020).

To use the information in pre-trained language models for NTMs, there have been several attempts in the most recent research of topic modeling. For example, (Hoyle et al., 2020) use knowledge distillation to combine NTM and BERT, they apply a BERT-based autoencoder as the teacher model to improve its NTM. CombinedTM of (Bianchi et al., 2021) concatenate the sparse BoW vector with the sentence embeddings extracted from the pre-trained BERT as the final input of an NTM. Despite the improved performance of NTMs, to our knowledge, existing works focus on the *shallow* combinations between NTMs and pre-trained lan-

guage models, i.e., they either use the contextual information of pre-trained language models as the input of NTMs or align the outputs of the two kinds of models. Therefore, there needs further study on how to build deep interactions between NTMs and pre-trained language models.

We in this paper propose a BERT-based Neural Topic Encoder (BNTE), a new encoder for NTMs to discover the latent topic distribution of a document, which deeply integrates with the transformer layers of BERT. Conventionally, the encoder of an NTM only considers the BoW data of a document as input and ignores the sequential orders of the words. Instead, the proposed BNTE encodes both the BoW data and the sequence of words of a document, which can be complementary to each other for learning a better topic distribution for the document. Specifically, BNTE first encodes a document's BoW data into the embeddings of a topic token by a neural network with multi-layer perception (MLP), which is expected to capture the topical information in the BoW data. The proposed encoder then inserts this topic token as a special token into the sequence of the words in the document. The new sequence with the topic token is then fed into a BERT encoder. With the multi-layer attention mechanism (Devlin et al., 2019) in BERT, the topic token mutually interacts with other words of the document in every layer of BERT, which makes the contextual information of BERT flow to the topic token and the topic information influences the learning of BERT. Finally, the embedding of the topic token in the last layer of BERT is used to generate the topic distribution of the document, which is further fed to the decoder of the NTM. Thanks to the in-depth integration with BERT, the proposed model achieves the state-of-art performances in multiple widely used evaluations of topic models, shown in the extensive experiments in the comparisons with many advanced models. There are several unique and appealing properties of our proposed model: *(1)* The proposed NTM encoder deeply interacts with the BERT encoder via the introduction of the topic token, which enables richer connections between the topic information and the contextual information of BERT. This is different from other models, which align/concatenate the input/output between BERT and NTMs. *(2)* Instead of using the output of a pretrained (fixed) BERT as the input of an NTM, our model jointly trains the NTM and the BERT parts. In this way, the two parts can help the learning of each other. *(3)* Our proposed encoder is compatible to many NTMs based on VAEs. Therefore, it is a better alternative to the existing encoder of NTMs, which enables them to use the contextual information of word sequences from BERT.

2 Background

2.1 Neural Topic Model

Given the target corpus of J documents: $M = \{m_j\}$, where the vocabulary contains V unique tokens. Let $\boldsymbol{x}_j \in \mathbb{N}^V$ denote the BoW vector of m_j, where x_{jv} represents the frequency of the v-th word in the document m_j. VAE-based NTMs assume that the topic proportion $\boldsymbol{z}$ is generated from a prior distribution $p(\boldsymbol{z})$, and $\boldsymbol{x}$ is generated from the conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$ that is usually modeled by a decoder. To perform inference on the model, NTMs also approximate the true intractable posterior $p(\boldsymbol{z}|\boldsymbol{x})$ with a neural encoder network that parameterizes the variational distribution $q(\boldsymbol{z}|\boldsymbol{x})$. NTMs are trained by maximizing the following Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{NTM}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log(p(\boldsymbol{x}|\boldsymbol{z}))] - \mathbb{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})), \quad (1)$$

The first expected log-likelihood term (or reconstruction term in VAEs) encourages that the variational posterior of the latent variables are good at explaning the data, while the second Kullback-Leibler (KL) divergence attempts to match the variational posterior over latent variable to its prior. Different NTMs are varying in the assumption of the prior distribution for topic proportion $\boldsymbol{z}$. For example, both ProdLDA (Srivastava and Sutton, 2017) and SCHOLAR (Card et al., 2018a) apply logistic norm as its prior, DVAE (Burkhardt and Kramer, 2019) shows that Dirichlet is another choice, while WHAI (Zhang et al., 2018) and Sawtooth (Duan et al., 2021) use Gamma distribution to enforce the sparsity of topic proportion.

2.2 Pre-trained Language Model

In the current paradigm of pre-trained models, methods like BERT and GPT have been shown to be effective for improving NLP tasks (Liu et al., 2019; Rogers et al., 2020). Those pre-trained models, usually have a fine-grained ability to capture the linguistic pattern, resulting in semantic token embeddings. Such contextual embeddings can be fed into downstream tasks as latent features

(Reimers and Gurevych, 2019; Bianchi et al., 2021), and usually achieve promising performance after slight fine-tuning.

Generally speaking, self-attention (SA) is the core idea behinds BERT. Given a sequence of N words $S := \{w_1, ..., w_N\}$, and its corresponding query matrix $\mathbf{Q} \in \mathbb{R}^{N\times d}$, key matrix $\mathbf{K} \in \mathbb{R}^{N\times d}$, and value matrix $\mathbf{V} \in \mathbb{R}^{N\times d}$, SA computes the matrix of output as :

$$\text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d})\mathbf{V}. \quad (2)$$

The self-attention means that we compute the dot products of the query with all keys, divided by $\sqrt{d}$, and then obtain the weights on the values after a softmax function. By stacking several SA layers, each word incorporates information from words that are semantic similar to it (via the normalized attention weight) layer by layer, resulting in the final embeddings. In practice, instead of performing a single SA, BERT uses Multi-Head Self-Attention (MHSA) to capture different semantic information from different latent subspaces. All parameters in BERT are trained via two unsupervised tasks: 1) Masked Language Model (MLM), and 2) Next Sentence Prediction (NSP). We refer readers to Vaswani et al. (2017) for more details.

3 Our Proposed Model

As discussed above, conventional NTMs usually take the sparse BoW vector as input, which focuses on extracting global semantic meaning of a document and however loses the local dependencies for ignoring the word orders. Different from them, natural language models, such as BERT, keep the ordering patterns, which are complementary to the BoW information modeled by NTMs. To model the contextual information of word sequences, we develop the BERT-based Neural Topic Encoder (BNTE), which encodes a document's BoW data into the embeddings of a topic token and uses BERT to capture the interactions between the topic token and other word tokens. Following the general framework of NTMs based on VAEs, the proposed encoder is responsible to learn the topic distribution of a document and a general decoder is applied to reconstruct the BoW vector of the document. Therefore, the proposed encoder is a better alternative to the existing NTM encoders and a plug-and-play module that is compatible to many NTMs. Without loss of generality, we use a recent NTM, Sawtooth (Duan et al., 2021), as an example to encompass our proposed encoder. We name the variant as Sawtooth-BNTE, which consists of two components as follows.

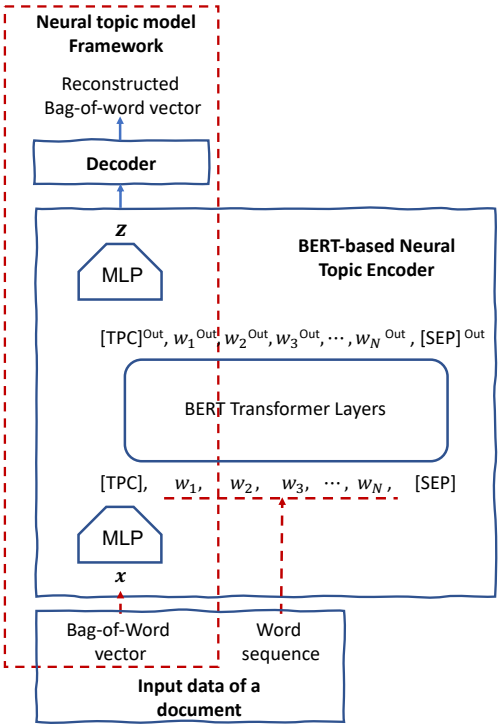

Figure 1: An overview of our model. The bottom part is the input including BoW and sequence of the document, middle is the BERT-based Neural Topic Encoder (BNTE), and upper is the decoder of NTM for reconstructing the BoW vector.

3.1 BERT-based Neural Topic Encoder

Motivated by the fact that the pre-trained BERT (Devlin et al., 2019) has been proved its effectiveness in a variety of downstream NLP tasks, we here adopt the pre-trained BERT as our text encoder. However, when the documents are longer than the length of input texts that BERT was designed for, a common phenomenon, it is unclear how exactly one would fine-tune BERT given existing documents. Given a document j, recall that BERT usually represents its input as a sequence, denoted as [[CLS], S_j, [SEP], where $S_j := \{w_1, ..., w_{N_j}\}$ includes the sequential tokens of the document. To adapt the pre-trained BERT to the document modeling naturally, we here replace the [CLS] in BERT with another special token $[\text{TPC}]_j$, which will be used for neural **topic** model. Specifically we express the input of BERT-based Neural Topic Encoder (BNTE) as:

$$\text{Input} := \{[\text{TPC}]_j, w_1, ..., w_{N_j}, [\text{SEP}]\}, \quad (3)$$

where j denotes the index of document. Following original BERT, for each token in Input, we can embed it into a vector $\boldsymbol{e}$ by combining the token

embedding (TE), segment embedding (SE) and the position embedding (PE): $\boldsymbol{e} = \text{TE} + \text{SE} + \text{PE}$. In terms of the special token $[\text{TPC}]_j$, we consider utilizing the document-level information to build its token embedding, which is specialized for each document m_j and formulated as

$$\text{TE}([\text{TPC}]_j) = \text{LayerNorm}(\text{MLP}(\boldsymbol{x}_j)), \quad (4)$$

where the V-dimensional BoW vector $\boldsymbol{x}_j$ is projected into the d-dimensional space by a single-layer network followed by the layer normalization. Moving beyond original BERT designs the token embedding for [CLS] with a one-hot vector shared by all documents, we assimilate the document-level information to realize the token embedding, which distinguishes $[\text{TPC}]_j$ from [CLS].

This simple combination attracts the following highlights: 1) The global BoW embedding carried by the $[\text{TPC}]_j$ and the sequential words embeddings can interact and improve each other due to the MHSA. In detail, the BoW information can attend to the contextual language knowledge within the sequential input, and in turn words have access to additional information about the document, which may be truncated and ignored because of the document-length limit in BERT; 2) Since the $[\text{TPC}]_j$ has the similar position embeddings and segment embeddings with [CLS], we can guarantee that our BNTE can reload the pre-trained weights from BERT; 3) By visualizing those words in document j that have high attention scores with $[\text{TPC}]_j$, one can explain and understand the latent semantic representation at each layer in topic model. Following other BERT-based models, we use the output embeddings of [TPC] as the document feature and formulate the encoding process as follows:

$$[\text{TPC}]_j^{\text{out}} = \text{BNTE}_\theta(\boldsymbol{x}_j, S_j), \quad (5)$$

where θ denotes the parameters in BNTE and $[\text{TPC}]_j^{\text{out}}$ is further adopted to infer the latent topic distribution as described below.

In terms of Sawtooth (Duan et al., 2021), it posits the Gamma distribution as the prior of the topic proportion $\boldsymbol{z}$. To infer $\boldsymbol{z}$, similar to the previous works (Zhang et al., 2018; Duan et al., 2021), we can adopt Weibull distribution to approximate its true posterior. The mainly reasons are that Weibull distribution resembles the Gamma distribution and there exists a simple reparameterization for $x \sim \text{Weibull}(k, \lambda)$, formulated as $x = \lambda(-\log(1-\epsilon))^{1/k}, \epsilon \sim \text{Uniform}(0, 1)$. Importantly, moving beyond the Sawtooth or other conventional NTMs, which infer topic proportion $\boldsymbol{z}$ using the BoW vector, we utilize the document-level contextual feature $[\text{TPC}]^{\text{out}}$ in (5). Formally, $q(\boldsymbol{z}|[\text{TPC}]^{\text{out}}) = \text{Weibull}(k_w([\text{TPC}]^{\text{out}}), \lambda_w([\text{TPC}]^{\text{out}}))$, where k_w and λ_w are two related neural networks parameterized by w. To sum up, as depicted in Fig. 1, we first embed the BoW vector $\boldsymbol{x}$ with BERT-based Neural Topic Encoder (BNTE), producing the document-level contextual feature $[\text{TPC}]^{\text{out}}$ at the position of the [TPC] token. Taking the $[\text{TPC}]^{\text{out}}$ as the input of inference network (MLP block in Fig. 1), we infer the topic proportion $\boldsymbol{z}$.

3.2 Decoder for Reconstructing BoW

Conditioning on the topic proportion $\boldsymbol{z}_j$ for document m_j, the decoder in Sawtooth aims to reconstruct the BoW vector $\boldsymbol{x}_j$ under the Poisson likelihood, expressed as:

$$\begin{aligned} \boldsymbol{x}_j &\sim \text{Pois}(\boldsymbol{\Phi}\boldsymbol{z}_j), \\ \boldsymbol{z}_j \sim \text{Gam}(r, c_j), \quad \boldsymbol{\Phi}_k &= \text{Softmax}(\boldsymbol{\rho}^T \boldsymbol{\alpha}_k), \end{aligned} \quad (6)$$

where $\boldsymbol{\Phi} \in \mathbb{R}_+^{V \times K}$ denotes the factor loading matrix (topic-term matrix), $\mathbb{R}_+ = \{x, x \geq 0\}$, K is the number of topics and V the vocabulary size. Besides, r and c_j are the shape and scale prior of the latent Gamma distribution; $\boldsymbol{\rho} \in \mathbb{R}^{D \times V}$ and $\boldsymbol{\alpha}_k \in \mathbb{R}^D$ denote the word embeddings and k-th topic embeddings, where D indicates the dimension of embedding space. The k-th topic $\boldsymbol{\Phi}_k \in \mathbb{R}_+^V$ is obtained by the inner product between the word embedding matrix $\boldsymbol{\rho}$ and corresponding topic embedding vector $\boldsymbol{\alpha}_k$, followed by the softmax normalization to enforce the simplex constrain in $\boldsymbol{\Phi}_k$. Rather than learning from the word co-occurrence alone, the topic distributions in Sawtooth are inferred by calculating the semantic similarities between topic and word embeddings, resulting in more coherent topics. Below we will omit the subscript j for simplicity.

3.3 Training Objective

The inference task is to learn the parameters of Sawtooth-BNTE, which can be summarized as $\boldsymbol{\Omega}$ by including the BNTE parameterized by θ, inference network parameterized by w, word embeddings $\boldsymbol{\rho}$, and topic embeddings $\boldsymbol{\alpha}$. Notably, we fine-tune the pre-trained BERT for adapting it to the corpora used in topic modelling. We can maximize the EBLO in Eq 1 to optimize the $\boldsymbol{\Omega}$. Our proposed Sawtooth-BNTE 1) allows the integration of

pre-trained linguistic knowledge into NTMs; 2) offers an efficient fine-tuning of BERT for corpora by replacing shared [CLS] token with the document-specific [TPC] and jointly learning with NTM. It also brings an additional benefit for alleviating the sentence length limitation of BERT, where BoW input provides the document-level information.

4 Related Work

This work develops a new topic model framework that combines NTMs with additional knowledge. NTMs are widely studied in many ways, for a review, we refer readers to (Zhao et al., 2021a).

One of the research directions of NTMs is to introduce metadata that is ignored by conventional topic models. Some previous works are proposed to address this issue. For example, SCHOLAR of Card et al. (2018b) combines the abilities of supervised LDA (SLDA, Blei and McAuliffe (2007)) and sparse additive generative models (SAGE, Eisenstein et al. (2011)), which provides a general algorithm for NTMs to incorporate word embeddings, document label and covariates with a variety of options. (Dieng et al., 2020; Duan et al., 2021) propose an embedding-based topic model (ETM), which directly models the similarity between words and topics in its generative process. After that, Sawtooth (Duan et al., 2021) is further proposed as an updated version of ETM.

In this paper, we aim to improve NTMs with the pre-trained language model and we choose Sawtooth as our base NTM although other choices are also available. Most recently, NTMs with pre-trained language models have obtained increasingly research interest. TopicBERT of Chaudhary et al. (2020) concatenates the outputs of NTMs and BERT directly to obtain a topic-aware document representation, they focus on supervised document classification not a pure document modeling framework. Hoyle et al. (2020) train a BERT-based autoencoder as a teach model (BAT), and use knowledge distilled from BAT to import its NTM. Bianchi et al. (2021) propose the Combined Topic Model (CombinedTM) to incorporate the pre-trained document contextualized representations from sentence BERT (SBERT, Reimers and Gurevych (2019)) into Product-of-Experts LDA (ProdLDA) of Srivastava and Sutton (2017) to improve the topic coherence. While CombinedTM is most related to our Sawtooth-BNTE, there are fundamental differences between CombinedTM and Sawtooth-BNTE in term of combination method, and base NTM. Specifically, CombinedTM incorporates the contextual embeddings by directly concatenating the BoW vector with the output embeddings of SBERT, while, in Sawtooth-BNTE, we replace the [CLS] token with [TPC] in the BERT, and naturally incorporate the BoW embeddings into the additional special token, so that the global BoW embeddings can interact with its words, resulting in more informative document representation. Moreover, Sawtooth-BNTE uses embedding representations of both words and topics, and the topic-term distributions are calculated by the semantic similarities between them, rather than being trained as a global parameter as in CombinedTM.

5 Experiments

5.1 Dataset

To evaluate and demonstrate the potential of the proposed model, we present a series of experiments on five commonly used dataset described as follows: *(1)* **AG** is a collection of more than 1 million news articles, gathered from more than 2000 news sources. We construct AG by choosing 4 largest classes from the original corpus (Zhang et al., 2015), where each class contains 30,000 training and 1,900 testing samples. *(2)* **DP** is constructed by picking 14 non-overlapping classes from DBPedia 2014 (Zhang et al., 2015). For each class, we here randomly choose 4,000 training samples and 1,000 testing samples. *(3)* **R8** is collected from Reuters 21,578 dataset, with 8 different categories. *(4)* **20NG** is a collection of approximately 20,000 newsgroup files, with the data organized into 20 different newsgroups, each corresponding to a different topic. *(5)* **Wiki103** is a version of WikiText dataset, which is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia. Considering R8 is already pre-processed, for other four datasets, we follow the pre-processing step in (Hoyle et al., 2020), where we tokenize and clean text by removing punctuation, standard stopwords, infrequent words steps. We use the default training/testing splits. Table 2 summarizes the statistics of the datasets.

5.2 Baselines

We consider the following baselines: *(1)* **SCHOLAR** (Card et al., 2018a), a VAE-based NTM that posits the logistic normal prior for

Model	AG			DP		
	Purity ↑	NMI ↑	ACC ↑	Purity ↑	NMI ↑	ACC ↑
DVAE	73.93 ±0.11	40.39 ±0.02	77.19 ±0.02	72.24 ±0.04	62.39 ±0.10	81.70 ±0.13
SCHOLAR	59.91 ±0.16	36.05 ±0.27	77.66 ±0.15	67.91 ±0.34	62.28 ±0.27	80.80 ±0.19
ETM	73.22 ±0.02	41.61 ±0.10	85.04 ±0.15	77.08 ±0.24	70.22 ±0.16	89.44 ±0.20
Sawtooth	70.86 ±0.11	43.11 ±0.12	84.57 ±0.13	72.77 ±0.20	71.15 ±0.12	89.61 ±0.19
BERT+Sawtooth	78.88 ±0.09	50.27±0.12	85.10±0.06	78.10 ±0.11	69.94 ±0.05	89.12 ±0.02
CombinedTM	79.69 ±0.08	51.28 ±0.15	85.92 ±0.07	79.25 ±0.17	71.15 ±0.14	**91.42** ±0.11
Sawtooth-BNTE (Ours)	**81.24** ±0.08	**51.97** ±0.24	**86.48** ±0.10	**81.23** ±0.11	**73.20** ± 0.22	90.25 ±0.08
	R8			**20NG**		
DVAE	82.13 ±0.20	50.74 ±0.26	90.54 ±0.17	35.95 ±0.14	29.18 ±0.11	47.77 ±0.04
SCHOLAR	80.50 ±0.14	45.92 ±0.20	87.85 ±0.07	47.26 ±0.16	39.74 ±0.20	52.18 ±0.18
ETM	77.52 ±0.20	46.14 ±0.08	93.79 ±0.22	45.96 ±0.09	38.11 ±0.11	54.72 ±0.06
Sawtooth	79.52 ±0.15	45.97 ±0.10	94.11 ±0.13	46.75 ±0.08	40.25 ±0.05	54.97 ±0.16
BERT+Sawtooth	82.62 ±0.08	49.87 ±0.09	94.53 ±0.10	45.26 ±0.04	40.10 ±0.12	56.24 ±0.08
CombinedTM	82.69 ±0.20	50.10 ±0.15	93.69 ± 0.17	48.25 ±0.16	47.54 ±0.25	56.83 ±0.11
Sawtooth-BNTE (Ours)	**83.95** ±0.06	**53.14** ± 0.18	**95.29** ±0.08	**48.32** ± 0.15	**47.88** ±0.16	**58.68** ±0.14

Table 1: Results on unsupervised document classification (measured with ACC) and clustering (measured with Purity and NMI) tasks. The best score of each dataset is highlight in bold.

Dataset	J	V	C	L
AG	127,600	11,347	4	20.12
DP	70,000	8,218	14	22.73
R8	7,674	7,688	8	65.72
20NG	18,864	2,000	20	107.38
Wiki103	269,503	2,000	N/A	124.77

Table 2: Summary statistics of our used corpora, where J is the number of documents, V the vocabulary size, C the number of classes and L the average length of the documents.

Method	NPMI ↑	WETC ↑	TD ↑
	AG		
DVAE	0.014	0.105	0.564
SCHOLAR	0.019	0.152	0.604
ETM	0.026	0.163	0.609
Sawtooth	0.024	0.176	**0.762**
BERT+Sawtooth	0.037	0.189	0.567
CombinedTM	0.044	0.187	0.705
Sawtooth-BNTE(Ours)	**0.057**	**0.201**	0.716
	DP		
DVAE	-0.003	0.070	0.492
SCHOLAR	0.012	0.105	0.604
ETM	0.025	0.134	0.697
Sawtooth	0.029	0.210	0.701
BERT+Sawtooth	0.028	0.176	0.652
CombinedTM	0.035	0.295	0.734
Sawtooth-BNTE(Ours)	**0.041**	**0.317**	**0.739**
	R8		
DVAE	-0.047	0.159	0.621
SCHOLAR	-0.055	0.142	0.634
ETM	-0.045	0.163	0.290
Sawtooth	-0.042	0.165	0.387
BERT+Sawtooth	-0.032	0.160	0.574
CombinedTM	-0.027	0.151	0.664
Sawtooth-BNTE(Ours)	**-0.018**	**0.170**	**0.676**
	20NG		
DVAE	-0.001	0.136	0.474
SCHOLAR	0.003	0.148	0.539
ETM	0.009	0.156	0.462
Sawtooth	0.010	0.132	0.504
BERT+Sawtooth	0.011	0.134	0.608
CombinedTM	0.015	0.144	0.619
Sawtooth-BNTE(Ours)	**0.024**	**0.167**	**0.705**
	Wiki103		
DVAE	0.052	0.225	0.388
SCHOLAR	0.046	0.184	0.525
ETM	0.054	0.264	**0.628**
Sawtooth	0.058	0.365	0.509
BERT+Sawtooth	0.062	0.374	0.558
CombinedTM	0.073	0.398	0.613
Sawtooth-BNTE(Ours)	**0.085**	**0.421**	0.604

Table 3: Topic quality of the learned topics on all datasets.

the topic proportion, and introduces metadata. SCHOLAR is equivalent to the ProdLDA (Srivastava and Sutton, 2017) when ignoring metadata and sparsity. *(2)* **DVAE** (Burkhardt and Kramer, 2019), another VAE-based NTM using a Dirichlet prior, whose reparameterization is implemented by rejection sampling variational inference. *(3)* **ETM** (Dieng et al., 2020), an embedded NTM that views words and topics as the trainable vectors living in the shared embedding space. *(4)* **Sawtooth** (Duan et al., 2021), another ETM-based NTM that uses the Gamma Belief Network (Zhou et al., 2015) as its decoder, which is adopted as our base NTM due to its promising performance in document modeling. *(5)* **BERT+Sawtooth**, a degraded version of our model, which uses BERT (Devlin et al., 2019) to extract contextual document embedding at the position of the [CLS] token and takes this embedding as input of the encoder of Sawtooth. Similar to our model, BERT+Sawtooth is optimized by maximizing the ELBO, where BERT is also fine-tuned. *(6)* **CombinedTM** (Bianchi et al., 2021), another BERT-based NTM that shares similar motivation with our proposed model. CombinedTM first feeds the sequential document into the pre-trained SBERT (Reimers and Gurevych, 2019) to

Methods	NPMI	Top-10 words
CombinedTM	-0.026	band, album, song, recording, harrison, had, recorded, songs, released, after
	-0.029	race, event, match, boat, championship, won, seconds, competition, wrestling, oxford
	0.108	film, role, films, script, cast, filming, production, reviews, bond, movie
	0.129	league, teams, stadium, club, football, team, cup, players, match, cricket
Sawtooth-BNTE(Ours)	0.128	album, albums, band, guitar, bands, billboard, charts, vocals, label, lyrics
	0.135	wrestling, championship, tournament, match, olympic, cup, champion, medal, hockey, race
	0.163	actress, film, films, actor, awards, comedy, opera, filming, drama, theatre
	0.128	cup, goals, league, scored, matches, liverpool, match, stadium, football, hockey

Table 4: Comparison of topics learned from CombinedTM and our proposed model on Wiki103, where the topics from latter are more clear and coherent than ones from former.

output its embedding, and then concatenate it with BoW as the input vector fed into ProdLDA.

5.3 Settings

We run all algorithms with the number of topics $K = 100$. For k_w and λ_w in the NTM-based encoder, we employ a fully-connected neural network with one hidden layer of 100 units and the SoftPlus as the activation function. The dimension of word and topic embeddings in ETM, Sawtooth and our proposed model is set as $D = 100$. For the pre-trained BERT, we use the uncased implementation of Huggingface [1], where d equals to 768. The optimisation of our model is done by Adam optimizer (Kingma and Ba, 2015) with learning rate 0.001 and batch size 6 for maximally 200 epochs. For baselines, we employ their official codes and default settings obtained from their official Github repositories.

5.4 Evaluation Metrics

A desired topic model is expected to extract representative topic proportion $\boldsymbol{z}$ of each document and the interpretable global topic-term matrix, i.e., $\boldsymbol{\Phi}$, where each column $\boldsymbol{\Phi}_k$ of $\boldsymbol{\Phi}$ corresponds to a topic and is a distribution over all tokens in the vocabulary. We here evaluate topic proportion and topic-term matrix for a comprehensive assessment.

Metrics about topic proportion $\boldsymbol{z}$. We consider measuring the performance of unsupervised classification and clustering tasks of different NTMs on AG, DP, R8 and 20NG, where the document labels are available. For document classification task, following (Qiang et al., 2020), we use a linear kernel Support Vector Machine (SVM) classifier and report the accuracy (ACC) on all datasets. For clustering task, we apply the K-Means algorithm and measure the clustering performance with Purity and NMI (Yin and Wang, 2014).

[1] https://huggingface.co/

Metrics about topic-term matrix $\boldsymbol{\Phi}$. To evaluate the learned topics, we adopt three criteria including Normalized Pointwise Mutual Information (NPMI), External Word Embeddings Topic Coherence (WETC) and Topic diversity (TD) (Bianchi et al., 2021). **NPMI** is calculated by the pointwise mutual information of each word pair over top-10 words of each topic. This requires the co-occurrence probability of any two words in a window size (e.g., 10), and the marginal probability of each word. We estimate these probabilities using empirical counts from additional dataset (e.g., Wikipedia), which has been proved to have a much higher correlation with human judgment than using the original text (Ding et al., 2018). Besides NPMI, we also adopt **WETC** to measure how semantic similar the key words in a topic are. Given a pre-trained word embedding matrix, such as GloVe (Pennington et al., 2014), we first compute the average pairwise cosine-similarity of the word embeddings of the top-10 words in a topic, and then average all the obtained scores over all topics to get the final WETC. The first two measures the topic coherence. Since the repeated topics may still have high topic coherence, we further report **TD**, which is defined as the percentage of the unique word in the top-25 words of all topics following (Dieng et al., 2020).

5.5 Quantitative Comparison

We run all experiments five times with different random seeds and calculate the mean and standard deviation. The unsupervised clustering and classification results are shown in Table. 1. We can find that **i**) Overall, our proposed model outperforms other NTMs in terms of both document clustering and classification tasks on almost all of the datasets, although with a slightly lower ACC on DP. This result reflects that our proposed model can extract more discriminative topic proportions to represent the input documents; **ii**) Although

BERT+Sawtooth introduces the pre-trained BERT into Sawtooth, it could not enhance the Sawtooth all the time. However, when compared with Sawtooth and BERT+Sawtooth, our proposed model improves both of them, by allowing the BoW information to interact with its sequential words via the special token [TPC], suggesting the validation of our proposed method; **iii**) Compared to the most related CombinedTM, our Sawtooth-BNTE provides better document representation on almost all of the datasets. We attribute this to the attention between BoW information and words in BNTE and also the joint training for BERT and NTM.

To measure the quality of the learned topics, we report the NPMI, WETC and TD of different models on all datasets in Table. 3. Generally, BERT-based NTMs, including BERT+Sawtooth, CombinedTM and our model, usually outperform other NTMs, which indicates the benefit of introducing contextual language knowlege offered by pre-trained BERT. And our proposed model produces the most coherent and diverse topics on almost all of the corpora, even with a slightly lower TD on AG and Wiki103 dataset. Especially, our proposed model provides very competitive results compared to its related work-CombinedTM. This is mainly because that CombinedTM simply concatenates BoW and the output embedding of [CLS] token in SBERT together as the input of encoder in NTM, while our proposed model enables richer connections between the BoW information and contextual word embeddings in BERT-based encoder. The rich linguistic knowledge embedded in pre-trained BERT and the efficient fine-tuning of BERT can complement the missing sequential information in BoW and help the learning of high-quality topics.

5.6 Exploring the [TPC] and Learned Topics

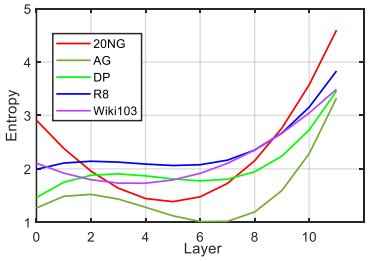

Figure 2: Attention entropy of [TPC] token at each layer over all used datasets.

As described in Sec. 2.2, one can access to the attention weights between [TPC] and its words at each head at each layer using Eq. 2. For [TPC], we first average all heads' attention at each layer, achieving the layer-level attention scores. To investigate how the [TPC] token attends to the sequential words, we then report the attention entropy at each layer in Fig. 2. Notably, the entropy measures the sparsity of attention. A higher entropy means that [TPC] attends to most words in the document, while a lower one denotes that [TPC] only attends to a small portion of the words in the document. We can see from Fig. 2 that the entropy of [TPC] shows a similar trend over all datasets: the BoW information carried by [TPC] token is inclined to attend to most words (coarse-grained) within the given document in the first few layers, and focus on some specific words (fine-grained) in the middle layers, and then gradually pay attention to most words (coarse-grained) in the last layers. This coarse-fine-coarse reading habit is somewhat similar to that of humans, and is beneficial for the proposed model to extract more informative topic proportions and explainable topics.

We also visualize four topics learned by CombineTM and our model, which are related to different domains, and report their NPMI scores in Table. 4. We observe that the topics discovered from ours are more cleanly focused on the query words: "band", "championship", "film", and "league". We attribute this to the above coarse-fine-coarse structure that can help the BoW vector incorporate language pattern from the sequential words.

6 Conclusion

To better discover latent topic distribution of a document, we developed a BERT-based Neural Topic Encoder (BNTE), a new encoder for NTMs, which deeply integrates with the BERT. BNTE encodes a document's BoW data into the embeddings of a topic token, and inserts it as a special token into the sequence of the words in the document, where the new sequence with the topic token is then fed into a BERT encoder. Thanks to the multi-layer attention mechanism in BERT, the topic token mutually interacts with other words of the document in every layer of BERT. By further mapping the topic token embed by BERT, we thus achieve more informative topic proportion. Extensive experiments demonstrate that our proposed model outperforms others on discovering high-quality topics and deriving better document representations. The proposed encoder can be an better alternative to the ones used in existing NTMs, which can be use to improve many existing methods.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. 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, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 759–766. Association for Computational Linguistics.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 121–128. Curran Associates, Inc.

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

Jordan L Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. *Applications of topic models*, volume 11. Now Publishers Incorporated.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018a. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018b. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2031–2040. Association for Computational Linguistics.

Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. 2020. Topicbert for energy efficient document classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1682–1690.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687*.

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

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1041–1048. Omnipress.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving neural topic models using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1752–1771. Association for Computational Linguistics.

Viet Huynh, He Zhao, and Dinh Phung. 2020. Otlda: A geometry-aware optimal transport approach for topic modeling. *Advances in Neural Information Processing Systems*, 33.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

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

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Open event extraction from online text using a generative adversarial network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 282–291. Association for Computational Linguistics.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. 2018. WHAI: weibull hybrid autoencoding inference for deep topic modeling. In *6th International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021a. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray L. Buntine. 2021b. Neural topic model via optimal transport. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mingyuan Zhou, Yulai Cong, and Bo Chen. 2015. The poisson gamma belief network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3043–3051.