Improving Neural Topic Models by Contrastive Learning with BERT

Anonymous ACL submission

Abstract

We present a general plug-and-play contrastive learning framework that improves existing neural topic models (NTMs) by incorporating the knowledge distilled from pre-trained language models. Recent NTMs have been applied to many applications and shown promising improvement on text analysis. However, they mainly focus on word-occurrences and are often optimized by maximizing the likelihood-based objective, which could lead to suboptimal topic coherence and document representation. To overcome the above bottleneck, we introduce an additional contrastive loss that pushes the topical representation of a document learned by an NTM close to the semantic representation of the document obtained from pre-trained language models. In this way, the prior knowledge of the pre-trained language models can enrich the contextual information of the target corpus for NTMs. Comprehensive experiments show that the proposed framework achieve the state-of-the-art performance. Importantly, our framework is general approach to improve most of the existing NTMs.

1 Introduction

A topic model (TM) discovers a set of interpretable topics from a target corpus, which can be used to derive topical representations of documents. TMs have been successfully applied in a wide range of applications such as document classification, keyphrase extraction, e-commerce recommendations, and clinical-admission analysis (Nan et al., 2019; Peinelt et al., 2020; Wang et al., 2019; Jin et al., 2018; Xu et al., 2018).

Recently, neural topic models (NTMs) (Miao et al., 2017; Srivastava and Sutton, 2017; Zhao et al., 2021b; Duan et al., 2021) have been a popular research direction in topic modeling due to their better flexibility and scalability than conventional TMs. Most of NTMs are inspired by the variational autoencoders (VAEs) (Kingma and Welling, 2014). Specifically, an NTM uses an encoder to derive the document's topic representation, indicating the topic proportion over the topics, and then feeds it into a decoder to reconstruct the document. Generally, the NTM is trained by maximizing the evidence lower bound (ELBO) of the likelihood of the observed documents. While, most of conventional TMs or NTMs purely learn from the statistical information of the target corpus, they may suffer from downgraded performance when there is less contextual information in the target corpus (Miao et al., 2017), such as short texts like tweets and news headlines. To tackle this issue, various approaches have been proposed, most of which use metadata such as pretrained word embeddings (Zhao et al., 2017; Inoue et al., 2021) and document labels (Card et al., 2018a) to complement to the contexts in the target corpus.

Recently, pre-trained language models (PLMs) such as BERT (Devlin et al., 2019; Reimers and Gurevych, 2019), and GPT (Brown et al., 2020) have been widely used various natural language processing (NLP) tasks. Trained on extremely large-scale corpora, PLMs can capture the semantic and syntactic information of natural languages. Given a document, a PLM can derive its semantic representation encoded by the CLS token (Adhikari et al., 2019). For the same document, the topical representation discovered from an NTM and the semantic representation discovered from a PLM are expected to be highly related. Therefore, PLMs can naturally serve as the sources of complementary contextual information for the training of NTMs.

Following this general idea, we introduce a new method that distills knowledge from PLMs to help learn NTMs better. The basic idea of our approach is very straightforward and intuitive: Given a PLM (pre-trained and fixed), we expect the topical representation of a document derived from a to-be-learned NTM to be close to the document's semantic representation derived from the PLM. In this

way, the prior knowledge of the PLM, pre-trained on large global corpora, can enrich the contextual information of the target corpus for learning NTM. To implement this idea, we are inspired by contrastive learning (CL) (Chopra et al., 2005; Robinson et al., 2021; Ma et al., 2021), which is a popular and successful self-learning approach originally proposed for image feature learning. In CL, the feature of an anchor image is expected to be closer to its positive samples' (e.g., the anchor image's augmentations) than its negative samples' (e.g., other images) (Saunshi et al., 2019; Tian et al., 2020). In our case, we propose a novel contrastive loss for training NTMs. Specifically, for an anchor document's topical representation learned by an NTM, we define its positive samples as the semantic representations generated from a PLM and push the topical representation close to them. Importantly, the proposed contrastive loss serves as an additional training objective to the original maximum likelihood estimation loss (e.g., ELBO) of existing NTMs without changing the model architectures of NTMs. That is to say, we propose a general plug-and-play technique that is flexible enough to improve on many existing NTMs. Our contributions are summarized as follows: ***(1)*** We propose a novel contrastive method that helps learn better NTMs by distilling knowledge from pre-trained language models, which tackles the issue of insufficient information in the target corpus for training NTMs. ***(2)*** The proposed approach is model agnostic and can be used to improve an arbitrary NTM. ***(3)*** Extensive experiments show that our proposed model achieves better document classification accuracy while discovering high-quality topics.

2 Background

2.1 Topic modeling and neural topic models

Suppose that $D = \{d_j\}$ is the corpora including J documents and we can represent each document d_j as a BoW count vector $\boldsymbol{x}_j \in \mathbb{N}^{V_t}$. Here V_t denotes the size of the vocabulary in topic modelling and x_{vj} is the number of times the v-th word occurs in the j-th document. In general, the goal of TM is to learn K shared topics $\{\boldsymbol{\phi}_k\}^{k=1:K}$ from the corpora and the representation $\boldsymbol{z}_j \in \mathbb{R}^K$ for document j, which indicates the document's topic proportions over K topics.

Most of existing NTMs follow the framework of VAEs (Kingma and Welling, 2014). A typical NTM consists of an encoder, which maps the BoW input $\boldsymbol{x}$ to its topic proportion $\boldsymbol{z}$, denoted as $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ for approximating the posterior $p(\boldsymbol{z}|\boldsymbol{x})$, and a decoder that generates $\boldsymbol{x}$ conditioned on the topic proportion $\boldsymbol{z}$, expressed as $p_\psi(\boldsymbol{x}|\boldsymbol{z})$, where we omit the subscript j for simplicity. Therefore, one can learn an NTM by maximising the ELBO of the marginal likelihood of BoW vector $\boldsymbol{x}$ in terms of θ, ψ, formulated as

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{x}, \theta, \psi) = \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log(p_\psi(\boldsymbol{x}|\boldsymbol{z}))] - \mathbb{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})), \quad (1)$$

where $p_\psi(\boldsymbol{x}|\boldsymbol{z})$ denotes the likelihood about the BoW vector $\boldsymbol{x}$ and the second term is the Kullback-Leibler (KL) divergence that regularises $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ to be close to its prior $p(\boldsymbol{z})$. Due to the unusable reparameterization trick in original VAEs for the commonly-used Dirichlet or gamma distributions in general TMs, various configurations of the prior distribution $p(\boldsymbol{z})$, data distribution $p_\psi(\boldsymbol{x}|\boldsymbol{z})$, posterior distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x})$, as well as different architectures of the decoder and encoder, have been developed for VAE-based NTM. We refer readers to Zhao et al. (2021a) for more details about NTMs.

2.2 Contrastive learning

Recent contrastive learning (CL) methods have been successfully applied in learning meaningful representations (van den Oord et al., 2018; Chen et al., 2020b). The main idea behind CL is that the more similar are two data points the closer they live in the latent space (Saunshi et al., 2019). Specifically, for an anchor data sample $\boldsymbol{x}$, one can find pairs of positive (similar) samples $(\boldsymbol{x}, \boldsymbol{x}^+)$ and negative (dissimilar) pairs $(\boldsymbol{x}, \boldsymbol{x}^-)$. The goal is to learn a function $f_\theta : \mathbb{R}^V \to \mathbb{R}^K$ that maps those associated samples $(\boldsymbol{x}, \boldsymbol{x}^+, \boldsymbol{x}^-)$ to the latent distribution $(\boldsymbol{z}, \boldsymbol{z}^+, \boldsymbol{z}^-)$. CL typically specifies the noise contrastive estimation (NCE) objective (Logeswaran and Lee, 2018):

$$\max_\theta \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}\left[\log \frac{\exp(\mathbf{z}\cdot\mathbf{z}^+)}{\exp(\mathbf{z}\cdot\mathbf{z}^+)+\beta\cdot\exp(\mathbf{z}\cdot\mathbf{z}^-)}\right], \quad (2)$$

where β is the strength of the constraint and $\beta = 1$ yields the usual form of the contrastive objective.

3 The proposed model

This paper proposes the Contrastive BERT-based framework to improve NTMs (CBTM for short) via distilling knowledge from PLMs, with the help of contrastive learning, whose overview is shown

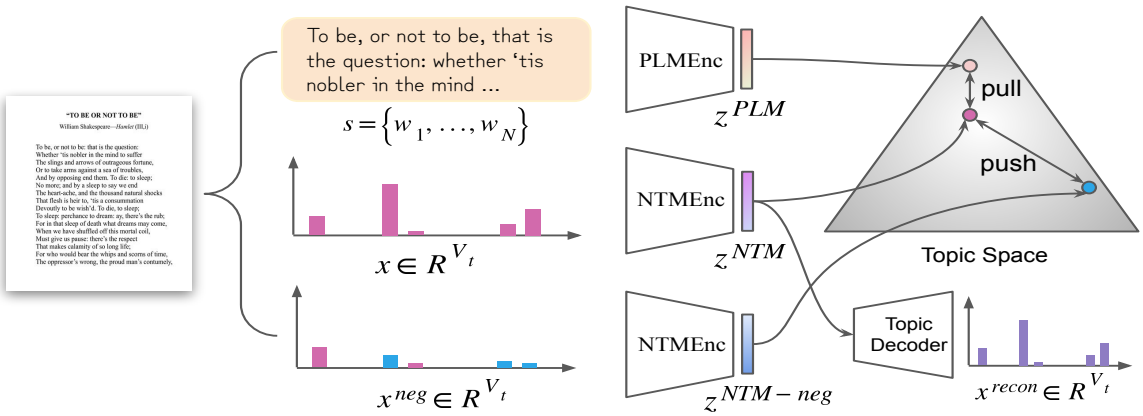

Figure 1: An overview of our proposed model. For a given document d, we can treat its BoW vector $\boldsymbol{x}$ as the anchor, where x-axis denotes the word index in the vocabulary and y-axis is the count value of x_v. Then we view the sequential representation $\boldsymbol{s}$ of document d as the positive sample of anchor $\boldsymbol{x}$ and construct the negative sample, where the blue bars in $\boldsymbol{x}^{\text{neg}}$ denote the perturbed weights of the selected words according to their tf-idf scores. Then, the samples $(\boldsymbol{x}, \boldsymbol{x}^{\text{neg}}, \boldsymbol{s})$ are fed into modality-specific encoders for inferring latent representation $(\boldsymbol{z}^{NTM}, \boldsymbol{z}^{\text{NTM-neg}}, \boldsymbol{z}^{\text{PLM}})$.

in Fig. 1. Given a document d, an NTM uses an encoder $q_\theta(\boldsymbol{z}^{\text{NTM}}|\boldsymbol{x})$ to embed d's topical representation $\boldsymbol{z}^{\text{NTM}} \in \mathbb{R}^K$ (i.e., topic proportion) from the BoW data of the document, i.e., $\boldsymbol{x} \in \mathbb{N}^{V_t}$. We denote $\boldsymbol{z}^{\text{NTM}} = \text{NTMEnc}(\boldsymbol{x})$. Note that we do not specify the implementation of the NTM used here and our method is expected to work with an arbitrary NTM. As discussed in Section 2.2, we consider document d's topic representation $\boldsymbol{z}^{\text{NTM}}$ as the anchor and the next key point is how to specify the positive and negative samples of the anchor. Although extensive study on the selection of positive and negative samples has been conducted in contrastive learning for image representation learning (Chen et al., 2020a,b), it has not been comprehensively investigated for documents or topic modelling. In this paper, we propose a novel selection strategy.

Selection of positive samples. For the given document d, we can also represent it as a sequence of words, denoted as $\boldsymbol{s} = \{w_1, \cdots, w_N\}$ where $w_n \in \{1 : V_l\}$ and V_l is the vocabulary size of PLMs. The semantic representation of the same document d can be obtained by using the CLS token from a PLM: $\boldsymbol{h} = \text{PLMEnc}(\boldsymbol{s})$. We are not limited to a specific PLM and theoretically any PLM that can discover semantic representations of documents are applicable in our framework, where we employ Sentence BERT (SBERT) (Reimers and Gurevych, 2019) without loss of generality.

Our basic idea is straightforward and intuitive. As $\boldsymbol{h}$ is extracted by a PLM, which is trained on large general corpora, we can consider that $\boldsymbol{h}$ is complementary to the orderless BoW information. To incorporate such knowledge into the NTMs, we can push $\boldsymbol{z}^{\text{NTM}}$ obtained from our to-be-learned NTM close to $\boldsymbol{h}$. Usually, $\boldsymbol{z}^{\text{NTM}} \in \mathbb{R}^K$ and $\boldsymbol{h} \in \mathbb{R}^O$ live in the different spaces, thus we first project $\boldsymbol{h}$ into a K-dimensional vector by introducing a learnable matrix $\mathbf{E} \in \mathbb{R}^{K \times O}$, expressed as $\hat{\boldsymbol{h}} = \mathbf{E}\boldsymbol{h}$. Since $\boldsymbol{z}^{\text{NTM}}$ is sampled from $q_\theta(\boldsymbol{z}^{\text{NTM}}|\boldsymbol{x})$, we also employ the variational inference network $q_w(\boldsymbol{z}|\hat{\boldsymbol{h}})$ to consider the uncertainty. Here, the parameter w is part of θ, which includes the parameters for mapping $\boldsymbol{x}$ into a K-dimensional vector and those for mapping the K-dimensional vector into parameters of posterior distribution (e.g., mean and variance in Gaussian distribution), where w belongs to the latter. That is to say, we only additionally introduce $\mathbf{E}$ as the to-be-learned matrix. We formulate $\boldsymbol{z}^{\text{PLM}} = \text{NTMEnc}(\hat{\boldsymbol{h}})$. In our case, for one document d, it is natural to use the $\boldsymbol{z}^{\text{PLM}}$ from the PLM as the positive sample of the topical representation $\boldsymbol{z}^{\text{NTM}}$ from the NTM, as both of them capture the semantics of a same document.

Selection of negative samples. Here, we follow a general principle of selecting negative samples for our framework: *A document's topical representation will change if the important words in the document are changed.* Recalling that we have $\boldsymbol{z}^{\text{NTM}} = \text{NTMEnc}(\boldsymbol{x})$, we can generate a new BoW data $\boldsymbol{x}^{\text{neg}}$ by perturbing the counts of the important words in $\boldsymbol{x}$ and have $\boldsymbol{z}^{\text{NTM-neg}} = \text{NTMEnc}(\boldsymbol{x}^{\text{neg}})$. It is natural to assume that $\boldsymbol{z}^{\text{NTM-neg}}$ should be different from $\boldsymbol{z}^{\text{NTM}}$ and we can use $\boldsymbol{z}^{\text{NTM-neg}}$ as the negative sample of $\boldsymbol{z}^{\text{NTM}}$. To generate $\boldsymbol{x}^{\text{neg}}$ from $\boldsymbol{x}$, we take the following steps: **1)** Following (Nguyen and Luu, 2021), we sort the words in document d by their tf-idf scores, and select top M tokens possess-

ing the highest tf-idf scores $\{w_1, \cdots, w_M\}$. It is reasonable to assume that these words mainly contribute to the topic of the document, i.e., they are relatively more important. **2)** We copy $\boldsymbol{x}$ to $\boldsymbol{x}^{\text{neg}}$, i.e., $\boldsymbol{x}^{\text{neg}} = \boldsymbol{x}$. **3)** With $\boldsymbol{z}^{\text{NTM}} = \text{NTMEnc}_\theta(\boldsymbol{x})$, we feed it into the NTM decoder to get the reconstructed weights of $\boldsymbol{x}$, denoted as $\boldsymbol{x}^{\text{recon-weight}}$, which is a normalized probability vector. We obtain the predicted BoW count vector, $\boldsymbol{x}^{\text{recon}}$ by $\boldsymbol{x}^{\text{recon}} = \boldsymbol{x}^{\text{recon-weight}} \sum_{v=1}^{V_t} \boldsymbol{s}$. **4)** Finally, we permute the weight of the selected top-M word in $\boldsymbol{x}^{\text{neg}}$ by $\boldsymbol{x}_m^{\text{neg}} = \boldsymbol{x}_m^{\text{recon}}$ for all $m \in \{1, \cdots, M\}$, where M is a hyperparameter of our framework.

Training NTMs by distilling from PLMs with contrastive loss. With the specification of the positive and negative samples, we introduce the following contrastive loss for training NTMs with PLMs:

$$\mathcal{L}^{\text{CL}}(\theta, \mathbf{E}) = \log \frac{\exp\left(\boldsymbol{z}^{\text{NTM}} \cdot \boldsymbol{z}^{\text{PLM}}\right)}{\exp\left(\boldsymbol{z}^{\text{NTM}} \cdot \boldsymbol{z}^{\text{PLM}}\right) + \exp\left(\boldsymbol{z}^{\text{NTM}} \cdot \boldsymbol{z}^{\text{NTM-neg}}\right)}, \tag{3}$$

where $\mathcal{L}^{\text{CL}}(\theta, \mathbf{E})$ is parameterized by θ and $\mathbf{E}$, θ is the parameter of the NTM encoder and $\mathbf{E}$ is the matrix for embedding the output of PLM to the K-dimensional vector.

Given a specific NTM, the proposed contrastive loss serves as an additional loss to the one used to train the NTM originally. Therefore, we aim to maximize the final objective, expressed as

$$\mathcal{L} = \mathcal{L}^{\text{NTM}}(\theta, \psi) + \lambda \mathcal{L}^{\text{CL}}(\theta, \mathbf{E}), \tag{4}$$

where ψ is the parameter of the decoder of the NTM and $\lambda > 0$ is a hyperparameter that controls the balance between the two losses. Note that for most of the popular NTMs, $\mathcal{L}^{\text{NTM}}(\theta, \psi)$ is the ELBO in (1), as introduced in Section 2.1.

As the introduced contrastive loss is independent to the original NTM loss and it only uses the output of the NTM encoder, our method can be used to improve an arbitrary NTM. That is to say our method is model agnostic. This will be comprehensively demonstrated in the experiments.

4 Related work

4.1 NTMs with pre-trained language models

The most closet work in NTMs to ours is utilizing pre-trained Transformer-based language models to improve the NTMs. For example, Bianchi et al. (2021) introduced a Combined Topic Model (CombinedTM) to incorporate the pre-trained document contextualized representations from SBERT Reimers and Gurevych (2019) into Product-of-Experts LDA (ProdLDA) of Srivastava and Sutton (2017) to improve the topic coherence. To improve the document-level understanding, Chaudhary et al. (2020) proposed TopicBERT by combining an NTM with a fine-tuned BERT, which concatenates the topic distribution and the learned BERT embedding of a document as the features for document classification. Hoyle et al. (2020) combined the advantages of these two approaches—the rich contextual language knowledge in pre-trained BERT and the intelligibility of NTMs—using knowledge distillation, which is denoted as BERT-based Autoencoder Teacher (BAT). The authors instantiated BAT to the two existing NTMs, including Scholar (Card et al., 2018b) (i.e. BAT+Scholar) and Wasserstein-LDA (Nan et al., 2019) (i.e. BAT+W-LDA).

4.2 Contrastive learning for texts

Originally proposed for images, contrastive learning also start to gain popularity in natural language processing tasks (Logeswaran and Lee, 2018; Xu et al., 2021). However, how contrastive learning helps in (neural) topic modeling has not been carefully studied. The most related work to ours is Contrastive Neural Topic Model (CNTM) [1] (Nguyen and Luu, 2021). Inspired by human behavior when comparing different documents, CNTM proposed a sampling strategy to construct positive and negative sample (i.e., BoW) and additionally introduced the contrastive objective to improve the NTMs. Although ours is alos on contrastive learning for NTMs, we have a different propose that is to distill knowledge from PLMs to help learn better NTMs. This propose leads to a different selection of the positive samples, which come from PLMs. For using external semantic knowledge extracted by PLMs, NTMs can be guided to better infer topical representation of documents as well as better topics.

5 Experiments

In this section, we study the performance of the proposed model and compare it to related NTMs on five benchmark textual data. As a desired TM should discover both accurate topic proportions and coherent topics, we consider topic interpretability

[1] We are unable to compare with CNTM (Nguyen and Luu, 2021) as their code is not publicly available.

and document classification, as described below.

5.1 Corpora

We run our experiments on five readily available datasets, which include regular and short documents and vary in scales, described as follows: ***(1)*** **20NG** consists of newsgroups including 18,846 articles evenly categorized into 20 different categories. The number of training samples is 11,314 and testing 7,532. ***(2)*** **Ohsumed** is a set of 13,929 unique cardiovascular diseases abstracts from MEDLINE, an on-line medical information database. The classification scheme consists of the 23 Medical Subject Headings (MeSH) categories of cardiovascular diseases group. After removing documents belonging to multiple categories, we obtain 3,357 documents in the training set and 4,043 documents in the test set. ***(3)*** **R8** and **R52** are two subsets of the Reuters 21,578 dataset. R8 contains 5,485 training and 2,189 test documents from 8 different classes. R52 consists of 6,532 training and 2,568 test documents and each of them is associated with 52 different labels. ***(4)*** **AG** News contains 496,835 categorized news articles from more than 2000 news sources. Following (Zhang et al., 2015), we choose the 4 largest classes from this corpus, using only the description fields. Each class contains 30,000 training samples and 1,900 testing samples, leading to 120,000 training samples and 7,600 testing samples.

For all datasets, we first clean and tokenize text following the preprocessing steps in (Yao et al., 2019). For the topic model, we additionally exclude standard stop words and low frequency words appearing less than 5 times. For 20NG and AG, we keep the 20,000 most frequent terms as the vocabulary. The statistics of the preprocessed datasets are summarized in Table 1.

Dataset	J_{all}	J_{train}	J_{test}	C	V_t	N
20NG	18,864	11,314	7,532	20	20,000	142.75
Ohsumed	7,400	3,357	4,043	23	14,157	135,82
R8	7,674	5,485	2,189	8	7,688	65.72
R52	9,100	6,532	2,568	52	8,892	69.82
AG	127,600	120,000	7,600	4	20,000	19.74

Table 1: Summary statistics of the datasets, where J denotes the number of documents, C the number of classes, V_t the vocabulary size of topic modelling and N the average length of documents in the corpus, respectively.

5.2 Baselines

To demonstrate the effectiveness of introducing cross-modality positive samples and PLM in improving the existing NTMs based on the contrastive loss, we consider several baselines for a fair comparison, including representative NTMs and NTMs with PLMs, described as follows: *1)* **ProdLDA** (Srivastava and Sutton, 2017) presents the effective auto-encoder variational Bayes (AEVB) based inference algorithm for LDA, and uses logistic normal distribution for the Dirichlet prior. *2)* **ETM** (Dieng et al., 2020), a VAE-based NTM, which assumes that words and topics live in the same embedding space, and draws each word from a categorical distribution whose natural parameter is calculated by the inner product between the embeddings of vocabulary and an embedding of its assigned topic. *3)* **Sawtooth** (Duan et al., 2021), a ETM-based hierarchical NTM, while, employs Poisson and Gamma distributions to model the BoW vector and latent topic proportion, respectively. *4)* **DVAE** (Burkhardt and Kramer, 2019), a VAE-based NTM with the Dirichlet prior, introduces rejection sampling variational inference for its reparameterization. *5)* **SCHOLAR+BAT** (Hoyle et al., 2020) uses DistillBERT as the teacher model and trains SCHOLAR in the knowledge distillation framework. Note that SCHOLAR is equivalent to the ProdLDA without metadata and sparsity. *6)* **CombinedTM** (Bianchi et al., 2021), a variant of ProdLDA that combines the BoW with the contextual document embeddings extracted from the pre-trained SBERT as input to produce more meaningful topics.

5.3 Settings

For all experiments, we set the number of topics K=100, the dimension of word and topic embeddings of ETM-based models d=100, and the batch size as 200. Below we select the recent Sawtooth (Duan et al., 2021) as the base NTM in our framework (CBTM-Saw) when comparing it with other baselines and also consider other NTMs in Section 5.6 for a comprehensively evaluations. We initialize word embeddings and topic embeddings from the Gaussian distribution $\mathcal{N}(0, 0.02)$. For topic encoder $q_\theta(\boldsymbol{z}|\boldsymbol{x})$, we employ an inference network stacked with a 3-layer V_t-256-100 fully-connected layer (V_t is the vocabulary size in topic modelling), followed by a softplus layer. Our framework is flexible for the choice of pre-trained language models and we here adopt the SBERT (Reimers and Gurevych, 2019), whose maximum length is 256 and output $\boldsymbol{h} \in \mathbb{R}^O$ is

a 768-dimensional vector. Here we introduce the embedding matrix $\mathbf{E} \in \mathbb{R}^{768\times100}$ to project $\boldsymbol{h}$ into the 100-dimensional vector $\hat{\boldsymbol{h}}$. We set $\lambda = 1$ and $M = 15$ for all experiments. We set the learning rate as 0.001 and dropout rate as 0.1. We train the proposed model for a maximum of 500 epochs using Adam optimizer (Kingma and Ba, 2015). For baseline models, we used default parameter settings as in their original papers or implementations. All experiments are performed on an Nvidia RTX 2080-Ti GPU and implemented with PyTorch.

5.4 Topic interpretability

Dataset	Method	TD ↑	TC↑	TS ↑
20NG	ProdLDA	**0.833**	-0.020	1.642
	ETM	0.238	-0.027	3.067
	Sawtooth	0.553	0.012	4.183
	DVAE	0.431	-0.024	2.147
	SCHOLAR+BAT	0.666	-0.008	3.143
	CombinedTM	0.741	-0.019	1.502
	CBTM-Saw(Ours)	0.639	**0.024**	**4.931**
R8	ProdLDA	**0.698**	-0.043	1.927
	ETM	0.137	-0.028	3.062
	Sawtooth	0.549	-0.026	7.593
	DVAE	0.154	-0.037	4.752
	SCHOLAR+BAT	0.448	-0.024	3.852
	CombinedTM	0.632	-0.021	1.785
	CBTM-Saw(Ours)	0.575	**-0.018**	**8.242**
R52	ProdLDA	0.627	-0.024	1.978
	ETM	0.180	-0.032	2.630
	Sawtooth	0.534	-0.011	6.985
	DVAE	0.152	-0.024	4.824
	SCHOLAR+BAT	0.553	-0.018	3.287
	CombinedTM	**0.630**	-0.012	1.784
	CBTM-Saw(Ours)	0.497	**-0.003**	**7.322**
Ohsumed	ProdLDA	**0.785**	-0.018	1.304
	ETM	0.226	0.038	2.71.
	Sawtooth	0.642	0.043	**7.005**
	DVAE	0.364	0.024	2.041
	SCHOLAR+BAT	0.706	0.043	3.036
	CombinedTM	0.711	0.040	1.267
	CBTM-Saw(Ours)	0.696	**0.052**	6.992
AG	ProdLDA	0.678	0.009	1.994
	ETM	0.306	0.125	2.951
	Sawtooth	0.527	0.207	7.752
	DVAE	0.519	0.218	0.647
	SCHOLAR+BAT	0.657	0.040	3.205
	CombinedTM	0.668	0.048	2.466
	CBTM-Saw(Ours)	**0.706**	**0.050**	**7.803**

Table 2: Topic quality over top 60% highest NPMI topics, where the best results are highlighted in boldface.

We comprehensively measure topic interpretability by blending three metrics: topic coherence (TC), topic diversity (TD) and topic specificity (TS). Given a reference corpus, **TC** measures the semantic relevance in the most significant words (top 10 words in our case) of a topic, which is computed by the Normalized Pointwise Mutual Information (NPMI) over the selected words of each topic (Dieng et al., 2020): $f(w_i, w_j) = \left[\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}\right] / \left[-\log p(w_i, w_j)\right]$, where $p(w_i, w_j)$ is the probability of words w_i and w_j co-occurring in a document and $p(w_i)$ is the marginal probability of word w_i, and both of them are estimated with empirical counts. Those models owing higher topic coherence are more interpretable topic models. As is implied by the name, **TD** measures how diverse the learned topics are. We define TD with the percentage of the unique word in the top 25 words of all topics (Zhao et al., 2020). TD that closes to 0 indicates redundant topics; that closes to 1 means more diverse topics. Besides TC and TD, we also report **TS**, which is used to measure how far a topic $p(\boldsymbol{\phi}_k|k)$ is from the overall distribution of words in the corpus $p(w)$. We calculate distance using KL divergence (Lee et al., 2021): $\text{TS} = \frac{1}{K}\sum_{k=1}^{K}\text{KL}(p(\boldsymbol{\phi}_k|k)||p(w))$. A larger distance means the distilled topics are more distinct; while a smaller distance suggests that the topics are more similar to the corpus distribution (overly general).

Since not all the learned topics are interpretable (Yang et al., 2015), we choose 60% topics with the highest NPMI, and report their average scores at Table 2. For the results, we have the following observations. Firstly, we can observe that topics discovered by our proposed model achieve the highest topic coherence (TC) across all corpora, while maintaining a competitive diversity (TD) and specificity (TS). This is because the PLM pre-trained from large general corpora, provides rich syntax and semantic information which can be incorporated as the complementary knowledge of the NTM with the contrastive loss. It is beneficial for inferring document's true topics in the scenarios where the BoW information is insufficient. Secondly, in terms of ProdLDA, while its topic diversity outperforms ours on a few datasets, it achieves low topic coherence and specificity, indicating its topics are diverse but less interpretable. Thirdly, compared to traditional NTMs, BERT-based NTMs including SCHOLAR+BAT and CombinedTM, usually produce more coherent and diverse topics. This result is in line with the previous study (Hoyle et al., 2020; Bianchi et al., 2021) that topic model itself can benefit from the general language knowledge of the pre-trained BERT. Among all the BERT-based NTMs, in general, our model performs the best.

Model	20NG	R8	R52	Ohsumed	AG
ProdLDA	58.42 ±0.24	89.26 ±0.17	80.14 ±0.09	41.85 ±0.24	79.96 ±0.20
DVAE	57.47 ±0.18	88.25 ±0.22	78.36 ±0.11	35.98 ±0.20	77.25 ±0.16
ETM	61.75 ±0.22	90.86 ±0.03	80.61 ±0.10	34.13 ±0.12	82.07 ±0.10
Sawtooth	64.65 ±0.21	92.60 ±0.11	80.92 ±0.05	42.51 ±0.09	83.04 ±0.09
SCHOLAR+BAT	66.03 ±0.08	92.98 ±0.24	82.17 ±0.08	44.20 ±0.10	85.06 ±0.19
CombinedTM	65.91±0.91	93.10±0.36	83.75 ±0.51	44.22±0.64	84.25 ±0.27
CBTM-Saw(Ours)	**66.46** ±0.10	**93.93** ±0.07	**84.35** ±0.07	**45.82** ±0.13	**86.25** ±0.11

Table 3: Test accuracy of different models on unsupervised document classification task. We run all methods 5 times and report the mean and standard deviation. The best scores of each dataset are highlighted in boldface.

5.5 Document classification

Considering doc-topic proportions can be viewed as unsupervised document representations, we perform document classification task and report accuracy (ACC) to evaluate the quality of such representation. Specifically, once we get the trained encoder network $q_\theta(\boldsymbol{z}|\boldsymbol{x})$, we feed the BoW vectors of testing documents into the encoder to collect the topic proportions. Then we apply logistic regression, which is trained on the proportions of training documents, to measure the classification performance of proportions of testing documents. Table. 3 summarizes the test accuracy of different NTMs in this task. As we can see, our proposed model obtains better classification performance than their baselines on all corpora, which confirms the effectiveness of our innovation of combining pre-trained language model and NTM in improving classification performance. Especially, even though SCHOLAR+BAT and CombinedTM incorporate the external knowledge learned by pre-trained language models into NTMs, both of them are inferior to our model. The main reason might be that we use the external information differently. In other words, moving beyond SCHOLAR+BAT and CombinedTM that incorporate the external language knowledge either through input or output in NTM, we build a general contrastive framework for NTM. It not only pulls together the positive pairs but also pushes away the negative samples, with the former borrowing the cross-modal language knowledge distilled from SBERT, and the latter bringing clearer classification boundaries, resulting in the SOTA accuracy.

5.6 Improving other NTMs

In previous experiments, we study the effectiveness of our proposed framework, where we adopt Sawtooth as the NTM. Since our framework is agnostic about the choice of the NTM, we in this experiment use other popular NTMs as the backbone of ours including ProdLDA (Srivastava and Sutton, 2017) and ETM (Dieng et al., 2020). Table. 4 shows the performance (topic quality and ACC) comparison between original NTMs and their improved variants in our framework on 20NG, R8, and Ohsumed. The performance of NTMs on three datasets has an improvement in most cases when combining our proposed contrastive framework, especially for the ETM. Although there are a slight decrease in TC on 20NG and TD on Ohsumed for ProdLDA when using our framework, our proposed models still achieve a better topic specificity and classification results. This observation validates the effectiveness of our proposed contrastive framework for enhancing existing NTMs. This suggests that our proposed plug-and-play framework can be flexibly used to enhance existing NTMs for topic modelling, without changing or re-designing the model architectures of NTMs on purpose, providing a simple but effective way for absorbing external semantic knowledge from PLMs.

Dataset	Method	TD	TC	TS	ACC
20NG	ProdLDA	0.833	**-0.020**	1.64	58.42
	Ours-ProdLDA	**0.860**	-0.032	**1.67**	**66.20**
	ETM	0.238	-0.027	3.067	61.75
	Ours-ETM	**0.263**	**-0.004**	**3.168**	**66.79**
R8	ProdLDA	0.698	-0.043	1.92	89.26
	Ours-ProdLDA	**0.723**	**-0.037**	**1.95**	**93.98**
	ETM	0.137	-0.028	3.06	90.86
	Ours-ETM	**0.149**	**-0.026**	**3.12**	**93.42**
Ohsumed	ProdLDA	**0.785**	-0.018	1.30	41.85
	Ours-ProdLDA	0.781	**-0.009**	**1.32**	**44.78**
	ETM	0.226	0.038	2.71	34.13
	Ours-ETM	**0.239**	**0.053**	**2.78**	**44.00**

Table 4: Performance of different models on 20NG, R8 and Ohsumed, respectively.

5.7 Ablation study and qualitative analysis

Number M of permuted tokens. To evaluate the impact of number M of permuted tokens in negative sampling, we report the performance of our proposed model on 20NG in Fig. 2 (a), where M is

Methods	NPMI	Top-10 words
Sawtooth	0.114	windows, software, pc, system, modem, dos, use, mac, unix, os
	-0.001	team, game, ca, season, play, hockey, roger, player, would, players
	-0.109	bike, dod, ride, motorcycle, riding, dog, bikes, helmet, bmw, nec
CombinedTM	0.003	windows, nt, microsoft, font, apps, os, type, fonts, seas, clarku
	-0.034	game, cup, go, series, goalie, playoffs, playoff, cmu, champs, beat
	-0.123	bike, riding, ride, bnr, helmet, bikes, mike, adobe, countersteering, hydro
CBTM-Saw(Ours)	0.128	mac, modem, port, apple, serial, card, sound, bit, pc, software
	0.024	baseball, game, edu, team, cubs, games, phillies, season, mets, braves
	0.006	bike, dod, ride, riding, motorcycle, bmw, bikes, helmet, motorcycles, behanna

Table 5: Learned topics of Sawtooth, CombinedTM, and CBTM-Saw(Ours) on 20NG dataset, where we choose three topics related to "software", "game" and "bike" query words.

ranging from 5 to 25. Besides, we further train two variants of CBTN-Saw (red lines) with the different schemes of the selection of positive and negative samples: without positive samples (green lines) and without negative samples (blue lines). We can find that 1) M can be selected to balance the document classification and topic quality. With tuning carefully for each dataset, one may get more better results than those reported in our experiments; 2) By combining the positive and negative samples together with contrastive loss, CBTM achieves better results than using either of them; 3) Compared with the negative samples, the positive samples generated from SBERT lead to more improvements, which is consistent with our motivation.

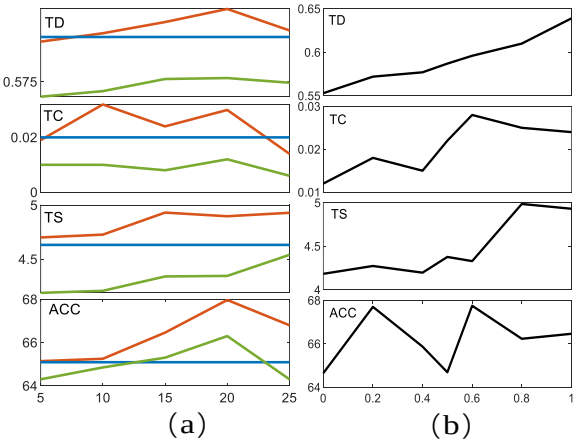

Figure 2: Shown in (a) and (b) are the ablation studies on 20NG about the number M of negative samples and the trade-off hyperparameter λ. In (a), CBTM-Saw and its two variants CBTM-Saw without positive samples, CBTM-Saw without negative samples are denoted as the red, green, and blue, respectively.

Trade-off hyperparameter λ. As shown in Fig. 2 (b), we further analyze the effect of trade-off hyperparameter λ which controls the weight of information incorporated from SBERT. Notably, we aim to explore the sensitiveness of our models for hyperparameter λ rather than exhaustively tuning this hyperparameter λ. We find that with the help of SBERT, the quality of the learned topics from CBTM have a greater improvement than only trained by NTM itself. We attribute this to the knowledge introduced from the pre-trained language model. Besides, the classification performance has a large accept range for λ, which means that CBTM is robust to document representation.

Visualization of learned topics. To investigate the effectiveness of our proposed model qualitatively, we visualize three topics related to query words including "software", "game" and "bike", which are extracted by Sawtooth, CombinedTM and our CBTM-Saw. For each topic, we select the top-10 words and report its NPMI at Table. 5. Compared with the Sawtooth and CombinedTM, the topics learned by our proposed CBTM-Saw are more coherent and explainable. This suggests that our proposed framework can enhance the learning of meaningful topics for assimilating document embeddings from PLM with contrastive loss.

6 Conclusions

We proposed a Contrastive learning framework called CBTM for neural topic models, which provides a straightforward but effective way for introducing semantic language pattern from pre-trained language models. For a document, CBTM views the document embeddings generated from pre-trained SBERT as the positive sampels, and permutes the weights of the key words as the negative samples. The additional contrastive loss pushes the latent distribution encoded from NTMs closer to the contextual representation distilled from BERT, while pulls away from the negative samples, resulting in more informative and distinguished latent distributions. Our model has shown appealing properties that are able to improve many existing NTMs without changing their model architectures.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: BERT for document classification. *CoRR*, abs/1904.08398.

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.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2031–2040. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 2031–2040.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

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.

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.

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)*.

Seiichi Inoue, Taichi Aida, Mamoru Komachi, and Manabu Asai. 2021. Modeling text using the continuous space topic model with pre-trained word embeddings. In *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, JUli 5-10, 2021*, pages 138–147. Association for Computational Linguistics.

Mingmin Jin, Xin Luo, Huiling Zhu, and Hankz Hankui Zhuo. 2018. Combining deep learning and topic modeling for review understanding in context-aware recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1605–1614.

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

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Moontae Lee, Sungjun Cho, Kun Dong, David Mimno, and David Bindel. 2021. On-the-fly rectification for robust large-vocabulary topic inference. In *International Conference on Machine Learning*, pages 6087–6097. PMLR.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xiaofei Ma, Cícero Nogueira dos Santos, and Andrew O. Arnold. 2021. Contrastive fine-tuning improves robustness for neural rankers. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 570–582. 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.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *arXiv preprint arXiv:2110.12764*.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.

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.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR.

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.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, and Shuming Shi. 2019. Topic-aware neural keyphrase generation for social media language. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2516–2526. Association for Computational Linguistics.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. 2018. Distilled wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1723–1732.

Peng Xu, Xinchi Chen, Xiaofei Ma, Zhiheng Huang, and Bing Xiang. 2021. Contrastive document representation learning with graph attention networks. *CoRR*, abs/2110.10778.

Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 308–317.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

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, Lan Du, Wray L. Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 635–644. IEEE Computer Society.

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 Buntine. 2020. Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537*.

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.