

# CWTM: Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling

Zheng Fang<sup>1</sup>, Yulan He<sup>1,2,3</sup> and Rob Procter<sup>1,3</sup>

<sup>1</sup>Department of Computer Science, University of Warwick, UK,

<sup>2</sup>Department of Informatics, King's College London, UK,

<sup>3</sup>The Alan Turing Institute, UK

{Z.Fang.4, Rob.Procter}@warwick.ac.uk

yulan.he@kcl.ac.uk

## Abstract

Most existing topic models rely on bag-of-words (BOW) representation, which limits their ability to capture word order information and leads to challenges with out-of-vocabulary (OOV) words in new documents. Contextualized word embeddings, however, show superiority in word sense disambiguation and effectively address the OOV issue. In this work, we introduce a novel neural topic model called the Contextualized Word Topic Model (CWTM), which integrates contextualized word embeddings from BERT. The model is capable of learning the topic vector of a document without BOW information. In addition, it can also derive the topic vectors for individual words within a document based on their contextualized word embeddings. Experiments across various datasets show that CWTM generates more coherent and meaningful topics compared to existing topic models, while also accommodating unseen words in newly encountered documents.

**Keywords:** Topic Modeling, BERT, Contextualized Word Embeddings

## 1. Introduction

Topic modeling has been widely used to explore latent themes within vast document collections, where each document is modeled as a mixture of topics, and a topic is represented by a list of words sorted by their co-occurrence association strength with the topic. Most existing topic models, whether probabilistic based or neural network based, rely on bag-of-words (BOW) information, either as input during training or as the target for training. BOW representations simply encode each document based on word occurrence within the document. It ignores syntactic and semantic relationships between words, consequently limiting the ability of topic models to produce more coherent topics.

While some topic models incorporate pre-trained word embeddings to capture syntactic and semantic information (Nguyen et al., 2015; Li et al., 2016; Zhao et al., 2017; Dieng et al., 2020; Gupta et al., 2019, 2020), they still fall short in capturing the contextual meaning of a word effectively, as the same words share an identical word embeddings irrespective of their surrounding context. Although some other topic models jointly learn topics and topic-specific word embeddings (Shi et al., 2017; Foulds, 2018; Zhu et al., 2020) to address this limitation, they do not leverage the advantages of pre-trained language models.

With the development of pre-trained language models like BERT (Devlin et al., 2018), T5 (Raffel et al., 2020), and GPT (Radford et al., 2019),

natural language processing (NLP) research has entered a new era. By capturing the surrounding context of each word, these models generate distinct contextualized word embeddings for each occurrence of a word in text. Contextualized word embeddings are superior to static word embeddings for word sense disambiguation, as they grasp the contextual nuances surrounding word usage. They also demonstrate improved capability in managing out-of-vocabulary (OOV) words (Akbik et al., 2018; Han and Eisenstein, 2019). Because contextualized word embeddings are generated contextually, they offer meaningful representations for rare or unseen words. Research has shown that they enhance the performance of various NLP tasks, including text classification, named entity recognition, and question answering (Devlin et al., 2018; Raffel et al., 2020; Radford et al., 2019).

Topic models can also leverage the advantages of pre-trained language models. The Contextualized Topic Model (CTM) (Bianchi et al., 2021) incorporates contextual information from BERT and achieves good performance. However, it only integrates contextual information from document embeddings and still relies on BOW representations as training inputs, thus missing out on the benefits of contextualized word embeddings and suffering from the OOV issue. Thompson and Mimno (2020), Grootendorst (2022) and Meng et al. (2022) also propose learning topics from pre-trained language models, but they treat topic modeling as a clustering task, which poses challenges in seamless integration with other downstream NLP tasks.

In our work, we developed a novel neural topic model combining contextualized word embeddings from BERT (Devlin et al., 2018). We assume that words with different semantic meanings have different topic information and show that by encoding topic information directly from contextualized word embeddings, our model can produce more coherent and meaningful topics, while also capable of handling unseen words from newly arrived documents. Specifically, we map each contextualized word embedding from BERT to a latent topic vector and aggregate these vectors to represent topics. The document-topic vectors are represented by applying weighted average pooling to the word-topic vectors in the same document.

In summary, our contributions are three-fold: **(1)** We have developed a novel neural topic model called the Contextualized Word Topic Model (CWTM), which integrates contextualized word embeddings from BERT, without relying on the BOW assumption. Our model demonstrates effective mitigation of the OOV issue. **(2)** We have compared our model with Latent Dirichlet Allocation (LDA) and other neural topic models across five datasets and shown that it can produce more coherent and meaningful topics. **(3)** We have demonstrated that the word-topic vectors learned from our model can improve the performance of downstream tasks such as named entity recognition, indicating that the vectors are semantically meaningful.<sup>1</sup>

## 2. Related Work

Early work has attempted to incorporate pre-trained word embeddings from GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) to capture the semantic and syntactic meaning in texts. Nguyen et al. (2015) integrated word embeddings into conventional topic models LDA (Blei et al., 2003) and GMM (Nigam et al., 2000) by replacing the topic-to-word Dirichlet multinomial component in these models by a two-component mixture of a Dirichlet multinomial component and the word embedding component. Li et al. (2016) combined GMM with word embeddings via the Generalized Po’lya Urn scheme to improve the performance of topic modeling for short texts. Zhao et al. (2017) introduced metaLDA incorporating various kinds of document and word meta information including word embeddings information. Gupta et al. (2019) proposed variants of neural autoregressive topic models called DocNADEe and iDocNADEe, using word embeddings as distributional priors, and in 2020 they took a step forward by extending the models to a lifelong neural topic modeling framework (Gupta et al., 2020).

---

<sup>1</sup>Our code can be found at <https://github.com/Fitz-like-coding/CWTM>.

With the development of large-scale, pre-trained language models, recent work has focused on combining contextualized representations from these models to capture word order information. Bianchi et al. (2021) introduced the contextualized topic model (CTM) that incorporates document embeddings from BERT to produce more coherent topics. They also introduced Cross-lingual Contextualized Topic Models with Zero-shot Learning to predict topics in different languages (Bianchi et al., 2020b). Mueller and Dredze (2021) extended the zero-shot CTM to fine-tune the contextualized document embeddings from the pre-trained language model and observed that it can facilitate cross-linguistic topic modeling. Similarly, Hoyle et al. (2020) adopted the knowledge distillation technique from Hinton et al. (2015) to combine neural topic models with the knowledge learned by the pre-trained language models.

While these studies have made good progress, they only make use of contextual information from document embeddings. None of them incorporates contextualized word embeddings into the topic modeling task. Moreover, they still require bag-of-words (BOW) information as training input or training target. Our work differs in that it incorporates contextualized word embeddings into the topic modeling process, without using any BOW information. Specifically, for each word in a document, our model encodes a word-topic vector from its contextualized word representation.

There has been some work that learns topics from pre-trained language models without using bag-of-words information (Thompson and Mimno, 2020; Grootendorst, 2022; Meng et al., 2022). However, they treat topic modeling as an embedding clustering task, and the topic distribution of each document learned does not follow the Dirichlet distribution assumption used in classical topic models, which is not the case for our model.

## 3. Contextualized Word Topic Model

We now introduce our Contextualized Word Topic Model (CWTM). Similar to the WLDA model from Nan et al. (2019), CWTM is also based on the Wasserstein autoencoder (WAE) (Tolstikhin et al., 2017). The WAE is an alternative framework to the variational autoencoder (VAE) (Kingma and Welling, 2013) for neural topic modeling. It consists of an encoder to map inputs to the latent space and a decoder to reconstruct inputs from the latent space. Similar to VAE, the WAE objective consists of a reconstruction term and a regularization term. However, the regularization term for WAE is designed to minimize the Wasserstein distance between the aggregated posterior and prior distribution, whereas the term for VAE is to minimize the

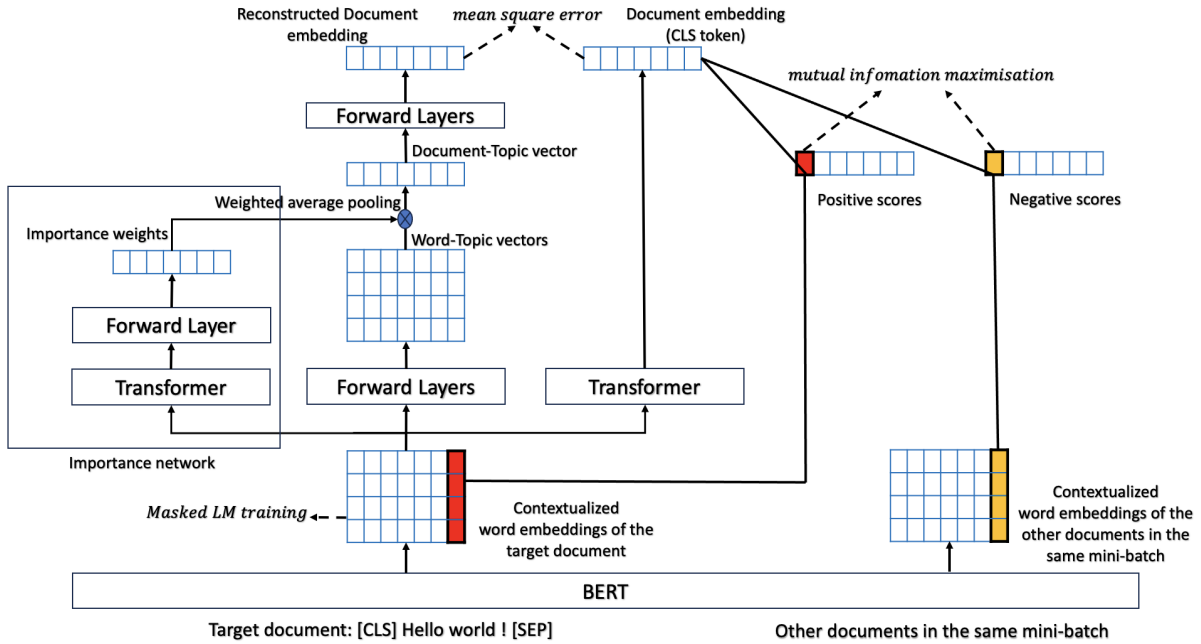


Figure 1: Model architecture. The contextualised word embeddings of the target document are encoded into word-topic vectors, and they are weighted average pooled to generate the document-topic vector, which is regularized to follow the Dirichlet distribution. The topic vector is then used to reconstruct the document embedding which was learned by a mutual information maximisation objective. A masked language model training objective is also added to regularize the word embeddings.

KL divergence between the posterior distribution and prior distribution. Nan et al. (2019) shown that WAE helps learn better topic representations than VAE-based topic models.

Different from the WLDA model which learns latent document-topic vectors from BOW representations, CWTM learns latent document-topic vectors from contextualized word embeddings. The encoder of CWTM consists of BERT and a multi-layer perceptron. Each word in a raw text document  $d$  is first converted to the contextualized word embedding  $e_w$  using BERT. The multi-layer perceptron then maps  $e_w$  to a lower-dimensional word-topic vector  $\theta_w$ . We also construct an importance network consisting of a transformer layer and a single-layer perceptron followed by the sigmoid activation to map each contextualized word embedding to the corresponding importance weight  $\alpha_w$ . The document-topic vector  $\theta_d$  is then represented as the weighted average pooling of these word-topic vectors. The decoder consists of a multi-layer perceptron mapping  $\theta_d$  back to the document embedding  $e_d$  of the document which is learned by a mutual information maximisation strategy. We present the model structure in Figure 1.

### 3.1. Combining Contextualized Word Embeddings

We start by considering learning latent document-topic vectors from document embeddings, which helps us understand the intuition behind combining contextualized word embeddings. We assume document embeddings contain the syntactic and semantic information required for topic information extraction. We could simply use the same WLDA model with document embeddings as input. Let  $e_d$  denote the document embedding of a raw text document  $d$ . The deterministic encoder from WLDA infers the latent document-topic vector:  $\theta_d = p(z|e_d) = \text{encoder}(e_d)$ , where  $z = [1, 2, \dots, Z]$  are topics and  $Z$  is the number of topics pre-specified. The decoder then reconstructs  $e_d$  from  $\theta_d$ . However, in this way, we cannot use decoder matrix weights to represent the topic words of each topic, as we don't have the vocabulary information. To overcome this, we consider learning topic information from contextualized word embeddings.

Let  $e_w$  denote the contextualized word embeddings from  $d$ , we can use a deterministic encoder followed by a softmax function to obtain the latent word-topic vector:  $\theta_w = p(z|e_w) = \text{encoder}(e_w)$ . Knowing  $p(z|e_w)$  we can then get the topic words of each topic by aggregating the topic-word vectors. We introduce the details in Section 4. We

use weighted average pooling over the latent topic-word vectors to get the document-topic vector  $\theta_d$ , as different words usually have different importance to the document:

$$\begin{aligned}\theta_d &= \sum_{w \in d} \beta_w \theta_w \\ \beta_w &= \frac{\alpha_w}{\sum_{\hat{w} \in d} \alpha_{\hat{w}}}\end{aligned}\quad (1)$$

where  $\alpha_w$  is the importance weight of contextualized word  $w$  in the document, which was learned by the importance network we introduced before. We enforce the document-topic vector to follow the Dirichlet distribution and introduce the details in Section 3.3. We then use the document-topic vector to reconstruct the document embedding of the document.

### 3.2. Learning Document Embedding

We utilise mutual information maximisation strategy (MI) (Kraskov et al., 2004) to learn the document embedding  $e_d$ . The MI maximizes the mutual information between  $e_d$  and the contextualized word embeddings  $e_w$  from the same document and minimizes its mutual information with the contextualized word embeddings  $e_{\bar{w}}$  from different documents.

We follow Zhang et al. (2020) to maximise the mutual information. Instead of using CNN layers to capture the n-gram local contextual dependencies of the input document as in Zhang et al. (2020), we add a transformer layer on top of BERT to capture the global contextual dependencies of the input document. We then use the embedding of the [CLS] token from the transformer to represent the document embedding  $e_d$ . Although there are different ways to represent document embeddings, such as mean pooling, we found no significant difference in the results. The document embedding and the contextualized word embeddings from BERT are then used for mutual information computation. The mutual information learning objective of the model is presented below:

$$L_{MI}(d) = -\log(f(e_w \cdot e_d)) - \log(1 - f(e_{\bar{w}} \cdot e_d)) \quad (2)$$

where  $f$  is the sigmoid activation and  $e_{\bar{w}}$  represents each contextualised word embedding from other documents in the same mini-batch.  $e_w \cdot e_d$  represents the dot product between them.

We further add the masked language model (MLM) objective  $L_{MLM}$  to regularise the contextualised word embeddings, ensuring that the contextualised information is preserved. The objective becomes:

$$L = L_{MI}(d) + L_{MLM}(d) \quad (3)$$

### 3.3. Distribution Matching

We enforce the document-topic vector to follow the Dirichlet distribution. We use the distribution matching function (Nan et al., 2019), enforcing  $\theta_d$  to follow Dirichlet distribution via equation 4:

$$\begin{aligned}\widehat{MMD}_{\mathbf{k}}(Q_{\Theta}, P_{\Theta}) &= \frac{1}{m(m-1)} \sum_{i \neq j} \mathbf{k}(\theta_i, \theta_j) + \\ &\frac{1}{m(m-1)} \sum_{i \neq j} \mathbf{k}(\theta_i', \theta_j') - \frac{2}{m^2} \sum_{i, j} \mathbf{k}(\theta_i, \theta_j')\end{aligned}\quad (4)$$

where  $\widehat{MMD}_{\mathbf{k}}(Q_{\Theta}, P_{\Theta})$  is the Maximum Mean Discrepancy (MMD) between the encoded distribution  $Q_{\theta}$  and the prior  $P_{\theta}$ . MMD is a distance measure between two distributions (Gretton et al., 2012). In our case,  $Q_{\theta} = \theta_d$  and  $P_{\theta}$  is the Dirichlet distribution.  $m$  is the size of the mini-batch and  $\mathbf{k}$  is the information diffusion kernel from Nan et al. (2019):

$$\mathbf{k}(\theta, \theta') = \exp(-\arccos^2(\sum_{z=1}^Z \sqrt{\theta_z \theta_z'})) \quad (5)$$

where  $Z$  is the number of topics.

Instead of only enforcing the document-topic vector to follow the Dirichlet distribution, we also enforce the topic-word vector  $\varphi_z$  to follow the Dirichlet distribution. The topic-word vector can be extracted by aggregating the topic vectors of words in the corpus. To reduce the training cost, we extract  $\varphi_z$  from each mini-batch and enforce it to follow Dirichlet distribution via  $\widehat{MMD}_{\mathbf{k}}(Q_{\Phi}, P_{\Phi})$ , where  $Q_{\Phi} = \varphi_z$  and  $P_{\Phi}$  is the Dirichlet distribution.

### 3.4. Final Objective

Finally, the decoder consisting of a multi-layer perceptron is used to reconstruct  $e_d$  from  $\theta_d$ , using the mean squared error (MSE) loss:

$$L_{REC}(d) = MSE(e_d, \text{decoder}(\theta_d)) \quad (6)$$

We jointly optimise all the objectives so the final objective function is:

$$\begin{aligned}L &= L_{MI}(d) + L_{MLM}(d) + L_{REC}(d) + \\ &\widehat{MMD}_{\mathbf{k}}(Q_{\Theta}, P_{\Theta}) + \widehat{MMD}_{\mathbf{k}}(Q_{\Phi}, P_{\Phi})\end{aligned}\quad (7)$$

To maximise the advantages of the pre-trained language model and retain the generalizability learned from pre-training, we also utilized the prompt learning strategy from Jiang et al. (2022). Prompts are trainable vectors that allow the optimisation of downstream tasks in an end-to-end manner. During the training, documents are fed into the frozen BERT with the pre-pended soft prompts.

## 4. Topic extraction

Instead of using the decoder matrix weights to extract the topic words of each topic, we extract the topic words of each topic by aggregating the topic vectors of words in the corpus. For example, for all the occurrences of the word “apple” in the corpus, we aggregate their topic vectors with regard to their importance weights in the corresponding documents to represent this word’s weight to each topic. We can then get the topic words of a topic by ranking all words based on the corresponding topic weight. The weight is guaranteed to be a positive value since we applied softmax activation in the encoder. This is similar to the way the Gibbs sampling LDA represents each topic. The difference is that Gibbs LDA assigns a fixed topic to each word, while CWTM assigns a soft topic to each word.

## 5. Experiments

In this section, we introduce the experimental setting of our work, including the datasets and the baseline models used. We focused on two aspects. First, whether our model can produce better quality topics than baseline models. Second, whether our model has a better ability to predict topics in new documents.

### 5.1. Datasets

Dataset	#train	#valid	#vocab	ave.doc.len	#class
20NG	13192	5654	20794	187.4	20
TagMyNews	22822	9782	12328	33.1	7
Twitter	3257	1421	1730	16.1	3
DBpedia	560000	70000	138708	46.1	14
AGNews	120000	7600	27521	37.8	4

Table 1: Dataset summary

We tested our model on five datasets, including the 20NG<sup>2</sup> dataset contains around 18k newsgroup posts on 20 topics; the TagMyNews<sup>3</sup> dataset contains around 32k short English news from 7 categories; the TweetEval emotion<sup>4</sup> dataset consists of 4672 tweets from three emotion classes, the DBpedia<sup>5</sup> dataset contains 560k documents from 14 ontology classes, and the AGNews<sup>6</sup> dataset contains 120k news articles from 4 categories. We summarize the basic statistics of these datasets in Table 1.

<sup>2</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\\_20newsgroups.html](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html)

<sup>3</sup><http://acube.di.unipi.it/tmn-dataset/>

<sup>4</sup>[https://huggingface.co/datasets/tweet\\_eval/viewer/emotion](https://huggingface.co/datasets/tweet_eval/viewer/emotion)

<sup>5</sup>[https://huggingface.co/datasets/dbpedia\\_14](https://huggingface.co/datasets/dbpedia_14)

<sup>6</sup>[https://huggingface.co/datasets/ag\\_news](https://huggingface.co/datasets/ag_news)

### 5.2. Baselines

We compared our model with four baselines, including LDA (Blei et al., 2003): A widely used topic model. We implemented the Gibbs Sampling LDA by ourselves<sup>7</sup>; ProLDA ((Srivastava and Sutton, 2017): a neural topic model uses a Gaussian approximation to replace the effect of the Dirichlet prior; WLDA (Nan et al., 2019): a neural topic model that utilizes Wasserstein distance to regularize the distribution matching; And CTM (Bianchi et al., 2021): a neural topic model utilizes sentence embeddings from pre-trained language models to enhance the quality of topics. All of these models require bag-of-words representations as the input, whereas CTM requires additional raw text documents as the input. We tuned the hyperparameters of each model and present the settings in Appendix A.

### 5.3. Evaluation Metrics

We used three metrics to evaluate each model, which are topic coherence, topic diversity and document classification accuracy. We cannot compute ELBO-based perplexity for WLDA and our model because they are not based on variational inference.

**Topic coherence** measures the semantic similarity between words within a given topic. It has been widely used in previous work (Srivastava and Sutton, 2017; Nan et al., 2019; Bianchi et al., 2021). We used the best performing measure C<sub>V</sub> from Röder et al. (2015), and used Wikipedia<sup>8</sup> as the external corpus. We focused on the top 10 words of each topic as in previous papers (Nan et al., 2019; Bianchi et al., 2021) and used the Palmetto library algorithm (Röder et al., 2015).

**Topic diversity** assesses the degree of diversity in a set of topics. It has been used as a supplementary measure to topic coherence. We computed the word embeddings-based centroid similarity<sup>9</sup> based on the top 10 words of each topic. The diversity score is expressed as 1 minus the similarity score. A higher score indicates more varied topics. For baseline models, we used BERT with the Aggregated Static Embeddings approach from Bommasani et al. (2020) to obtain the static embedding for each topic word. We used the contextualized word embeddings from the first layer of BERT and used mean pooling to distil the static embeddings. Bommasani et al. (2020) found this generally performs better than pre-trained static word embeddings. This also allows us to obtain

<sup>7</sup>we will share the code upon paper acceptance

<sup>8</sup>[https://hobbitdata.informatik.uni-leipzig.de/homes/mroeder/palmetto/Wikipedia\\_bd.zip](https://hobbitdata.informatik.uni-leipzig.de/homes/mroeder/palmetto/Wikipedia_bd.zip)

<sup>9</sup>We used the code from OCTIS: <https://github.com/MIND-Lab/OCTIS/tree/master>

		LDA	ProdLDA	WLDA	CTM	CWTM
20NG	Coherence	0.550±0.008	0.394±0.010	0.371±0.014	0.393±0.012	<b>0.600±0.020</b>
	Diversity	0.630±0.004	0.593±0.014	0.594±0.003	0.604±0.009	<b>0.664±0.005</b>
TagMyNews	Coherence	0.586±0.002	0.560±0.009	0.435±0.021	0.508±0.011	<b>0.626±0.016</b>
	Diversity	0.615±0.003	0.620±0.002	0.589±0.003	0.616±0.002	<b>0.676±0.005</b>
Twitter	Coherence	0.333±0.010	0.318±0.001	0.310±0.020	0.342±0.012	<b>0.357±0.005</b>
	Diversity	0.495±0.002	0.523±0.005	0.334±0.023	0.514±0.004	<b>0.58±0.0020</b>
DBpedia	Coherence	<b>0.571±0.012</b>	0.538±0.020	0.555±0.037	0.477±0.014	0.569±0.011
	Diversity	0.649±0.006	0.677±0.004	0.627±0.001	0.655±0.004	<b>0.692±0.004</b>
AGNews	Coherence	0.578±0.003	0.550±0.013	0.463±0.025	0.510±0.003	<b>0.633±0.020</b>
	Diversity	0.617±0.002	0.623±0.005	0.575±0.008	0.617±0.007	<b>0.680±0.004</b>

Table 2: Topic coherence and topic diversity results for 50 topics. We ran each experiment three times and report the average results. We also report the standard deviation.

the embeddings for rare words in the corpus. For CWTM, since it encodes word-topic vectors from contextualized word embeddings, different occurrences of the same word have different weights for each topic. For a fair comparison, a weighted average of the contextualized word embeddings from all occurrences of a word is used to represent its word embeddings for each topic. We also used contextualized word embeddings from the first layer of BERT to maintain consistency with the baseline models.

**Document classification** has also been used in many previous works to assess the predictive performance of the latent document-topic vectors (Li et al., 2016; Gui et al., 2019; Nan et al., 2019). Higher accuracy indicates better predictive performance. We applied a logistic regression classifier with a default parameters setting. We used the document-topic vectors from the testing set as the input and conducted five-fold cross-validation.

#### 5.4. Benchmark Results

We present the topic coherence and topic diversity results of using 50 topics in table 2. It can be observed that CWTM achieved competitive performance compared with baseline models on both topic coherence and topic diversity measures. It achieved the best coherence and diversity scores on the 20NG, TagMyNews, Twitter, and AGNews datasets. It also has the second-best coherence score and best diversity score on the DBpedia dataset. It suggests that encoding topic information from contextualized word embeddings can help generate more coherent and meaningful topics. Although CTM incorporated contextual document embeddings, it did not show significant advantages over other baseline models. One possible reason is that the sentence embedding model used by CTM can only accept up to 128 tokens in the input document, so it cannot handle long documents well.

We also plot the performance of different models

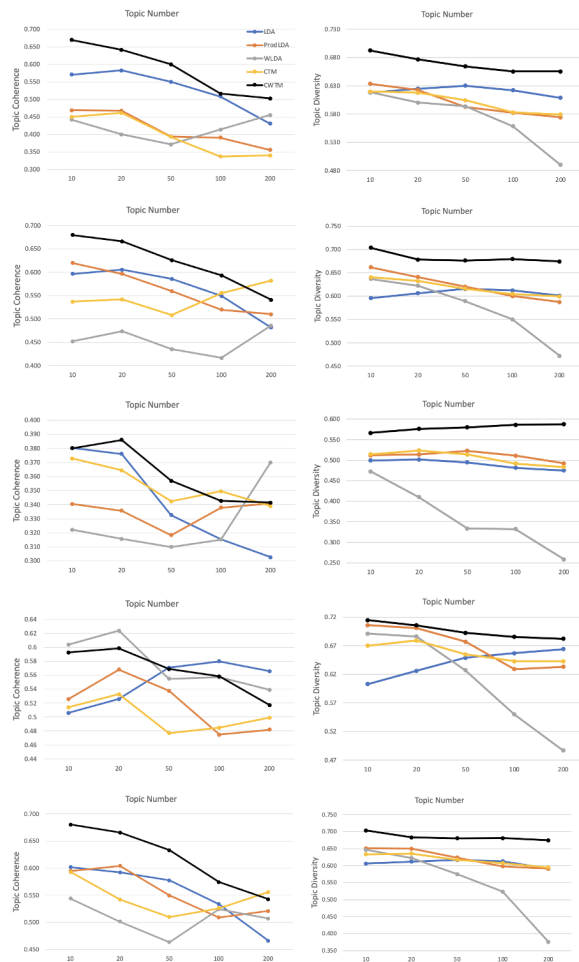


Figure 2: Topic coherence and topic diversity scores across different numbers of topics by different models for 20NG (top row); TagMyNews (2nd row), Twitter (3rd row), DBpedia (4th row), and AGNews (bottom row).

with different numbers of topics settings in Figure 2. We set the number of topics  $Z = \{10, 20, 50, 100, 200\}$ . It can be observed that overall

		CWTM	w/o $L_{MI}$	w/o $L_{MLM}$	w/o $L_{REC}$	w/o $\widehat{MMD}_k(Q_\Theta, P_\Theta)$	w/o $\widehat{MMD}_k(Q_\Phi, P_\Phi)$	w/o Importance
20NG	Coherence	<b>0.641±0.021</b>	0.396±0.038	0.349±0.028	0.514±0.037	0.519±0.034	0.596±0.005	0.517±0.027
	Diversity	0.677±0.011	0.618±0.009	0.000±0.000	0.655±0.024	<b>0.690±0.015</b>	0.639±0.009	0.607±0.002
	Classification	<b>0.530±0.024</b>	0.206±0.043	0.058±0.000	0.352±0.048	0.511±0.014	0.512±0.027	0.526±0.012
TagMyNews	Coherence	<b>0.666±0.018</b>	0.566±0.019	0.421±0.002	0.525±0.043	0.597±0.030	0.653±0.011	0.606±0.021
	Diversity	0.678±0.007	0.662±0.015	0.000±0.000	0.663±0.006	<b>0.690±0.006</b>	0.639±0.006	0.626±0.002
	Classification	0.819±0.001	0.576±0.034	0.252±0.000	0.626±0.071	0.810±0.003	0.820±0.001	<b>0.828±0.011</b>
Twitter	Coherence	<b>0.386±0.010</b>	0.332±0.020	0.361±0.035	0.378±0.014	0.355±0.005	0.378±0.017	0.368±0.006
	Diversity	0.576±0.004	0.553±0.029	0.021±0.025	0.573±0.008	<b>0.592±0.008</b>	0.541±0.009	0.564±0.007
	Classification	<b>0.641±0.013</b>	0.472±0.059	0.394±0.001	0.593±0.027	0.610±0.034	0.627±0.018	0.606±0.020
Dbpeida	Coherence	<b>0.599±0.015</b>	0.391±0.033	0.315±0.009	0.519±0.109	0.513±0.015	0.529±0.017	0.509±0.013
	Diversity	0.705±0.004	0.643±0.016	0.000±0.000	0.669±0.011	<b>0.718±0.002</b>	0.651±0.007	0.647±0.025
	Classification	0.947±0.004	0.619±0.029	0.079±0.000	0.816±0.115	0.947±0.013	<b>0.962±0.002</b>	0.950±0.009
AGNews	Coherence	<b>0.666±0.014</b>	0.620±0.010	0.390±0.023	0.600±0.032	0.512±0.026	0.661±0.006	0.567±0.010
	Diversity	0.683±0.003	0.675±0.005	0.000±0.000	0.648±0.016	<b>0.693±0.012</b>	0.650±0.006	0.632±0.009
	Classification	0.866±0.005	0.745±0.038	0.255±0.000	0.784±0.031	0.867±0.004	0.859±0.005	<b>0.874±0.002</b>

Table 3: Ablation study of CWTM. CWTM represents our model with all components included.

CWTM consistently outperforms other models in terms of the topic coherence scores on the 20NG, TagMyNews, Twitter, and AGNews datasets. It also has the second-best coherence scores on the DBpedia dataset when the number of topics is below 100. We can also conclude from the figure that CWTM produces more diverse topics than other baseline models across all numbers of topics settings. We also found that WLDA performed poorly on the topic diversity measure. It’s diversity score dropped dramatically when we increased the number of topics.

	20NG	TagMyNews	Twitter	DBpedia	AGNews
LDA	<b>0.589±0.004</b>	0.783±0.009	0.513±0.009	0.864±0.007	0.858±0.002
ProdLDA	0.359±0.010	0.681±0.003	0.437±0.008	0.769±0.003	0.762±0.008
WLDA	0.426±0.016	0.747±0.007	0.415±0.019	0.907±0.008	0.824±0.014
CTM	0.494±0.033	0.783±0.009	0.570±0.021	0.881±0.003	0.864±0.002
CWTM	0.530±0.024	<b>0.819±0.001</b>	<b>0.641±0.013</b>	<b>0.947±0.004</b>	<b>0.866±0.005</b>

Table 4: Document classification results for different models

We also present the document classification results in Table 4. We set the number of topics to 20. It can be observed that CWTM outperforms existing models on four datasets. Only LDA has better classification performance on the 20NG dataset. The excellent classification performance indicates that CWTM can produce document-topic vectors with good predictive power. We also plot the classification performance of different models with different numbers of topics settings in Appendix B. The results are consistent with Table 4.

## 5.5. Ablation Study

We study the effect of different objective functions and the importance network in this section. We present the performance of our model without each objective function and the importance network in Table 3. We set the number of topics to 20. It can be observed that with all components included, CWTM achieved the best topic coherence scores over all datasets. By excluding the mutual information loss ( $L_{MI}$ ), masked language model loss ( $L_{MLM}$ ), or reconstruction loss ( $L_{REC}$ ),

CWTM’s performance on all three evaluation metrics dropped significantly, suggesting that these objectives are critical to the overall performance of the model. By excluding the document level distribution matching objective ( $\widehat{MMD}_k(Q_\Theta, P_\Theta)$ ), CWTM produced slightly more diverse but much less coherent topics, indicating that this objective is beneficial for learning coherent topics, while slightly limits the model to learn more diverse topics. Without the topic level distribution matching objective ( $\widehat{MMD}_k(Q_\Phi, P_\Phi)$ ), CWTM’s topic coherence and diversity scores on all datasets dropped, suggesting that this objective is essential to learn coherent and diverse topics. To exclude the importance network, we treat words in a document as having equal weights, i.e.  $\alpha_w = 1$ . We can observe that without the importance network, the coherence and diversity scores dropped significantly, proving that different words have different contributions to a document. We present the topic words of CWTM on different datasets in Appendix C.

## 5.6. Handling Out-of-Vocabulary Words

	20NG	TagMyNews	Twitter	DBpedia	AGNews
ave.vocab_Test1	29.4	16.0	4.9	17.6	18.0
ave.vocab_Test2	56.7	15.1	7.5	20.5	18.2
ave.new_vocab_Test1	1.9	0.9	0.1	1.2	1.0
ave.new_vocab_Test2	27.6	5.9	3.6	8.2	7.0
#Test1	5990	6151	720	1180	4589
#Test2	1252	7264	1924	8192	3783

Table 5: The vocabulary difference between the Test1 and Test2 datasets. “ave.vocab” represents the average vocabulary size of each document. “ave.new\_vocab” represents the average unobserved vocabulary size in each document. #test1 and #test2 represent the number of documents in each test set.

We now describe how we compared different models’ performance in handling OOV words. For each corpus, we first randomly sampled 1000 documents as the training set. We then sampled two testing sets Test1 and Test2. For each document

		LDA	ProdLDA	WLDA	CTM	CWTM
20NG	$\geq 90\%$	0.415±0.013	0.139±0.008	0.166±0.009	0.359±0.023	0.371±0.017
	$\leq 70\%$	0.293±0.008	0.166±0.006	0.160±0.008	0.310±0.026	0.374±0.020
	diff	0.122±0.010	-0.027±0.01	0.006±0.013	0.050±0.013	<b>-0.004±0.01</b>
TagMyNews	$\geq 90\%$	0.708±0.010	0.512±0.011	0.462±0.029	0.718±0.019	0.804±0.008
	$\leq 70\%$	0.537±0.009	0.307±0.006	0.33±0.015	0.601±0.025	0.721±0.012
	diff	0.171±0.003	0.206±0.011	0.133±0.022	0.116±0.01	<b>0.083±0.006</b>
Twitter	$\geq 90\%$	0.489±0.022	0.466±0.013	0.421±0.016	0.55±0.023	0.595±0.021
	$\leq 70\%$	0.455±0.010	0.443±0.007	0.437±0.011	0.507±0.021	0.589±0.015
	diff	0.034±0.031	0.023±0.020	-0.016±0.015	0.043±0.018	<b>0.006±0.008</b>
Dbpeida	$\geq 90\%$	0.838±0.011	0.678±0.017	0.363±0.043	0.859±0.007	0.827±0.025
	$\leq 70\%$	0.800±0.015	0.580±0.012	0.294±0.039	0.839±0.011	0.837±0.020
	diff	0.037±0.017	0.098±0.025	0.069±0.028	0.021±0.01	<b>-0.01±0.023</b>
AGNews	$\geq 90\%$	0.826±0.006	0.642±0.007	0.527±0.027	0.831±0.008	0.871±0.004
	$\leq 70\%$	0.732±0.005	0.479±0.010	0.475±0.03	0.777±0.008	0.85±0.006
	diff	0.093±0.007	0.163±0.015	0.051±0.053	0.054±0.015	<b>0.021±0.005</b>

Table 6: Document classification accuracy over Test1 and Test2. “ $\geq 90\%$ ” means each document in Test1 has over 90% of the vocabulary words observed in the training set. “ $\leq 70\%$ ” means each document in Test2 has less than 70% vocabulary words observed in the training set. “diff” shows the accuracy difference between Test1 and Test2. The closer the difference is to 0, the better the performance of handling OOV words. We ran each model 5 times.

in Test1, over 90% of the vocabulary words were observed in the training set, and for each document in Test2, less than 70% of the vocabulary words were observed in the training set. In other words, documents in Test2 have more OOV words than documents in Test1. We set the number of topics to 20 and trained different models on the training set. We fed the document-topic vectors from the testing sets to a logistic regression classifier and conducted five-fold cross-validation. A model insensitive to OOV words is expected to have consistent classification performance over the two different testing sets. We summarize the basic statistics of the sampled Test1 and Test2 sets in Table 5.

We report the results in Table 6. We observed that CWTM achieved more consistent results over the different testing sets than baseline models. The accuracy difference between CWTM on Test1 and Test2 is smaller than that of the baseline models on all datasets. Student t-tests on accuracy difference show that CWTM outperforms other baselines on almost all datasets, except the WLDA model on the 20NG ( $t=-1.29$ ,  $p=0.12$ ) and AGNews ( $t=-1.28$ ,  $p=0.12$ ) datasets, ProdLDA model on the Twitter dataset ( $t=-1.13$ ,  $p=0.14$ ). However, the higher classification accuracy of CWTM suggests that CWTM is more competitive. This proves that incorporating contextualized word embeddings into topic models can help alleviate the OOV issue.

## 5.7. Word-topic Vector Quality

CWTM can learn latent word-topic vectors for words in a document. To evaluate the predictive performance of the latent word-topic vectors, we used the Named Entity Recognition (NER) metric. We trained CWTM on the conll2003 NER dataset (Sang and De Meulder, 2003) and concatenated the latent vectors with the contextualised word embeddings from BERT’s last layer and fed them into a single-layer perceptron for the conll2003 NER task. To learn the latent word-topic vectors using CWTM, we set the latent vector size to 50, the epoch number to 30 and the learning rate to  $1e-4$ . To train the single-layer perceptron, we set the epoch number to 10 and the learning rate to  $2e-3$ . We froze the parameters of BERT and CWTM when feeding the vectors. We used BERT as the baseline and fine-tuned it within 2 epochs using MLM objective on the target corpus. We ran the experiment five times and report the average results.

	Precision	Recall	F1
BERT	0.792	0.839	0.814
BERT+CWTM	0.807	0.850	0.828

Table 7: NER performance. The first row shows the results of using only contextualised word embeddings from BERT as the input. The second row shows the results of using concatenated vectors as the input.

The results are shown in Table 7. It can be seen that by adding word-topic vectors we can obtain better results for the NER task. A student t-test on



f1 scores ( $t=-3.75$ ,  $p=0.003$ ) shows this difference was statistically significant. This means that the latent word-topic vectors learned from CWTM are semantically meaningful.

## 6. Conclusions

We have developed a novel neural topic model called CWTM. The model uses the raw text document as input and infers topic information directly from the contextualized word embeddings of the document. In addition to learning the latent document-topic vector of a document, it also learns the latent word-topic vector of each word in the document. Experimental results show that overall CWTM outperforms the baseline models in topic coherence, topic diversity and document classification measures, which means that CWTM can produce more coherent and meaningful topics. The experiments on the OOV words prove that CWTM is more insensitive to unseen words from newly arrived documents.

Experiments with the CoNLL2003 NER dataset further demonstrates the benefits of word-topic vectors learned from CWTM and shows that CWTM can provide new input information for downstream NLP tasks. In the future, we aim to conduct human evaluations to assess the performance of the model on real-world tasks.

## Acknowledgements

ZF receives the PhD studentship jointly funded by the University of Warwick and China Scholarship Council. YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/2, EP/V020579/2).

## Bibliographical References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.
- 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 (Volume 2: Short Papers), pages 759–766, Online. Association for Computational Linguistics.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.
- BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.
- BSI. 1973b. *Natural fibre twines*. BS 2570, British Standards Institution, London. 3rd. edn.
- A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.
- J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.
- N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 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*.
- Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.
- James Foulds. 2018. Mixed membership word embeddings for computational social science. In *International Conference on Artificial Intelligence and Statistics*, pages 86–95. PMLR.

- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. 2019. Neural topic model with reinforcement learning. 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)*, pages 3478–3483.
- Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. 2019. Document informed neural autoregressive topic models with distributional prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6505–6512.
- Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schuetze. 2020. Neural topic modeling with continual lifelong learning. In *International Conference on Machine Learning*, pages 3907–3917. PMLR.
- Xiaochuang Han and Jacob Eisenstein. 2019. Un-supervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv preprint arXiv:1904.02817*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.
- Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.
- Alexander Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving neural topic models using knowledge distillation. *arXiv preprint arXiv:2010.02377*.
- Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.
- Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3021–3035.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.
- Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152.
- Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Aaron Mueller and Mark Dredze. 2021. Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. *arXiv preprint arXiv:2104.05064*.
- Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.
- Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.
- Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39:103–134.
- 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.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.
- Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. 2017. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 375–384.
- Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.
- Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).
- S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.
- Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, pages 33–45. Springer.
- Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626*.
- Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.
- Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*.
- He Zhao, Lan Du, Wray Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644. IEEE.
- Lixing Zhu, Yulan He, and Deyu Zhou. 2020. A neural generative model for joint learning topics and topic-specific word embeddings. *Transactions of the Association for Computational Linguistics*, 8:471–485.

## Appendix A: Parameters Settings

We trained all models using the training data and report the results for the test data. We used full-length texts as input. To transform documents into BoW representations, each dataset was properly pre-processed with word lemmatizing for each word, and all stop words were removed. We also removed all numeric and non-alphabetic characters, as well as frequently occurring words in each corpus, i.e., the top 10 most frequently occurring words in each corpus. When training CWTM, we did not preprocess the documents because it accepts raw text documents as input. When presenting the topics, we preprocessed the topic words generated by CWTM for a fair comparison with other models.

For Gibbs sampling LDA, we set  $\alpha = 1/t$  and  $\beta = 1/t$ , where  $t$  is the number of topics for the model and found it outperforms other settings. We trained it over 2000 iterations. For prodLDA and WLDA, we set the batch number to 256 and trained them over 500 epochs for the 20NG, TagMynews and Twitter datasets. We trained them over 10 epochs for the Dbpedia dataset since the number of documents in Dbpedia is much larger. We found no significant difference in the results compared to training them over 7-9 epochs. We used an ADAM optimizer with high momentum  $\beta_1 = 0.99$  and a learning rate of 0.002 as it does in Nan et al. (2019). We set the *keep\_prob* parameter of ProdLDA to {0.0, 0.2, 0.4, 0.6} and only report the best result from them. We set the Dirichlet prior of WLDA to 0.1 which was suggested by the original paper. The Dirichlet noise of WLDA is set to {0.0, 0.2, 0.4, 0.6} and we report the best result from them. For CTM, we used the best-performing sentence transformer “all-mpnet-base-v2”<sup>10</sup> to convert each document into a document embedding. We trained it over 100 epochs for the 20NG, TagMynews and Twitter datasets and 2 epochs for the Dbpedia dataset. We tuned the *n\_sample* parameter and found that setting it to 40 gives the best results. For CWTM, we built our model on top of “BERT-base-uncased” and we set the epoch number to 20 for the 20NG, TagMynews and Twitter datasets and set it to 1 for the DBpedia dataset. The Dirichlet prior was set to 0.1 and the batch size was set to 16. The learning rate was set to 1e-3. We also applied a linear scheduler with 10% warmup steps for CWTM. For the soft prompt we used, we set the length of the prompt to 10. We trained each model three times to report the average results. All experiences were conducted with NVIDIA GeForce RTX 3090.

## Appendix B: Document Classification Results

We plot the classification performance of different models with different numbers of topics settings in Figure B1. The number of topics is set to  $Z = \{10, 20, 50, 100, 200\}$ . CWTM achieved better performance than the other models on the TagMyNews, Twitter, Dbpedia, and AGNews datasets. Although LDA outperforms CWTM on the 20NG dataset, CWTM has better performance than the other neural topic model baselines.

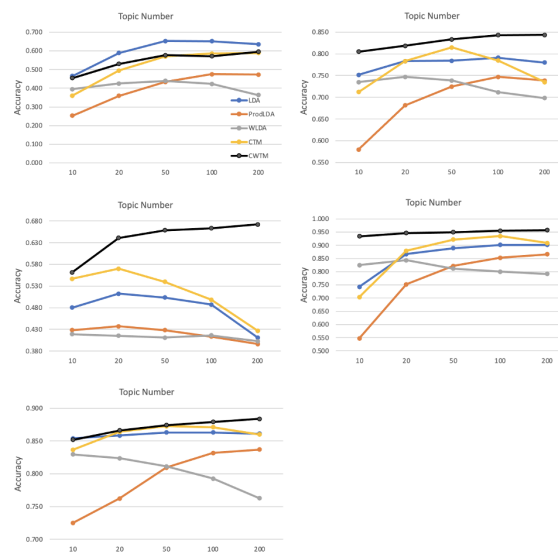


Figure B1: Document classification accuracy across different numbers of topics by different models for 20NG (top row left); TagMyNews (top row right), Twitter (2nd row left), DBpedia (2nd row right), and AGNews (bottom row).

<sup>10</sup>[https://www.sbert.net/docs/pretrained\\_models.html](https://www.sbert.net/docs/pretrained_models.html)

## Appendix C: Topic Words from CWTM on Different Datasets

Topic	Top 10 words
topic 1	key, chip, radio, circuit, phone, power, battery, encryption, signal, line
topic 2	car, bike, engine, mile, brake, tire, wheel, rear, dealer, bmw
topic 3	god, christian, bible, jesus, scripture, christ, church, sin, lord, holy
topic 4	space, nasa, orbit, doctor, patient, disease, medical, launch, drug, shuttle
topic 5	game, player, hockey, team, baseball, playoff, pitcher, nhl, play, goal
topic 6	drive, card, driver, disk, bus, memory, ram, controller, mac, machine
topic 7	argument, agree, moral, opinion, discussion, question, point, assume, belief, fact
topic 8	armenian, gun, child, fbi, weapon, police, arab, government, law, jew
topic 9	window, program, application, post, unix, code, software, command, server, run
topic 10	test, david, ditto, stuff, cheer, hey, deleted, tony, sex, michael

Table 8: Topic words from CWTM on the 20NG dataset.

Topic	Top 10 words
topic 1	nuclear, earthquake, tsunami, plant, tornado, quake, radiation, power, oil, reactor
topic 2	stock, bank, billion, investor, market, fund, price, financial, share, debt
topic 3	yankee, inning, sox, phillies, mets, run, hitter, pitcher, jay, baseball
topic 4	broadway, theater, musical, idol, tony, revival, play, show, stage, star
topic 5	study, cancer, risk, drug, diabetes, patient, disease, woman, heart, treatment
topic 6	open, round, federer, murray, nadal, french, andy, djokovic, final, master
topic 7	trial, court, case, police, accused, charge, prosecutor, judge, charged, guilty
topic 8	apple, google, microsoft, android, tablet, phone, ipad, browser, service, window
topic 9	connecticut, east, kentucky, ncaa, butler, ohio, duke, notre, big, texas
topic 10	rebel, libyan, force, libya, gaddafi, nato, government, japan, military, syrian

Table 9: Topic words from CWTM on the TagMyNews dataset.

Topic	Top 10 words
topic 1	customer, twitter, ur, tweet, email, account, post, awe, service, app
topic 2	fucking, fuck, shit, guy, hilarious, hell, pout, man, funny, team
topic 3	terrorism, anger, outrage, terror, black, rage, pakistan, angry, fuming, trump
topic 4	amazing, blue, lively, news, music, rock, musically, tonight, great, free
topic 5	hilarious, horrible, fuming, fury, depressing, rage, furious, bitter, irritate, nightmare
topic 6	good, happy, lost, year, work, back, night, home, tomorrow, wait
topic 7	depression, fear, sadness, anxiety, afraid, terrible, bad, worst, hate, cry
topic 8	today, phone, guy, year, work, service, back, thing, im, call
topic 9	live, broadcast, ly, gbbo, lol, birthday, gt, snap, wow, snapchat
topic 10	watch, musically, blue, horror, nightmare, absolutely, delight, exhilarating, glee, playing

Table 10: Topic words from CWTM on the Twitter dataset.

Topic	Top 10 words
topic 1	aircraft, car, engine, wing, seat, locomotive, fighter, model, motorcycle, air
topic 2	journal, university, newspaper, published, editor, peer, reviewed, academic, college, magazine
topic 3	high, public, grade, student, pennsylvania, township, secondary, education, historic, house
topic 4	football, played, footballer, player, play, professional, league, hockey, playing, team
topic 5	directed, starring, drama, book, movie, comedy, star, story, written, thriller
topic 6	navy, ship, class, launched, hm, commissioned, submarine, laid, vessel, built
topic 7	band, record, studio, rock, song, released, singer, track, single, ep
topic 8	mountain, peak, river, range, lake, mount, alp, elevation, summit, flow
topic 9	plant, found, genus, moth, endemic, habitat, flowering, native, tree, grows
topic 10	persian, iran, romanized, rural, gmina, poland, population, province, village, voivodeship

Table 11: Topic words from CWTM on the Dbpedia dataset.

Topic	Top 10 words
topic 1	stock, price, dollar, market, investor, higher, future, high, rise, rose
topic 2	court, charge, case, police, arrested, charged, lawsuit, judge, trial, drug
topic 3	oil, space, moon, spacecraft, crude, nasa, titan, station, saturn, planet
topic 4	arsenal, goal, liverpool, chelsea, united, manchester, striker, england, test, champion
topic 5	corp, million, billion, deal, buy, bid, group, business, takeover, share
topic 6	inning, sox, yankee, league, giant, astros, dodger, cub, twin, nl
topic 7	music, game, store, video, dvd, nintendo, sony, make, song, ipod
topic 8	baghdad, militant, iraqi, iraq, insurgent, rebel, killed, soldier, attack, troop
topic 9	microsoft, linux, software, version, window, source, server, system, application, enterprise
topic 10	open, round, final, championship, master, cup, champion, federer, seed, hewitt

Table 12: Topic words from CWTM on the AGNews dataset.