Coordinated Topic Modeling

Pritom Saha Akash **Jie Huang** **Kevin Chen-Chuan Chang**
University of Illinois at Urbana-Champaign, USA
{pakash2, jeffhj, kcchang}@illinois.edu

Abstract

We propose a new problem called coordinated topic modeling that imitates human behavior while describing a text corpus. It considers a set of well-defined topics like the axes of a semantic space with a reference representation. It then uses the axes to model a corpus for easily understandable representation. This new task helps represent a corpus more interpretably by reusing existing knowledge and benefits the corpora comparison task. We design ECTM, an embedding-based coordinated topic model that effectively uses the reference representation to capture the target corpus-specific aspects while maintaining each topic's global semantics. In ECTM, we introduce the topic- and document-level supervision with a self-training mechanism to solve the problem. Finally, extensive experiments on multiple domains show the superiority of our model over other baselines.[1]

1 Introduction

We often ask questions about well-defined topics when we read articles (e.g., news/research articles). E.g., in news domain, such a question can be like "what is in *entertainment* today? (it's about *oscar 2022*)" or in academic domain, it can be like "what is trending in *machine learning* in 2016? (it's about *deep learning*)". This is a practical human trait while understanding information in a text corpus. Rather than finding arbitrary topics, people often want to explore the text based on some well-defined topics. E.g., in news domain, such topics are *business*, *politics*, *sports*, etc.

A well-defined topic is not merely a name (i.e., surface name); it generally also has a representation (e.g., word distribution) that can be obtained from a large text corpus, which we may call the *reference representation*. E.g., in such representation of the topic *sports*, words like *play*, *game*, and *player* are mainly dominated. Similarly, for *politics*, these are *vote*, *election*, *party*, etc. However, a topic's representation generally deviates from its reference to other corpora. E.g., while describing a news corpus specific to the USA, it is likely to use words such as *nfl*, *baseball* or *football* to represent *sports*, while words like *cricket*, *ipl* or *run* are more relevant for India.

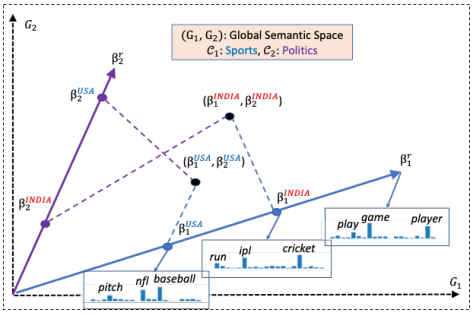

Figure 1: Coordinated interpretation of Topic model

The above scenario can be explained by the concept of the coordinate system in geometry. Specifically, we can consider a set of well-defined topics as the axes or basis of semantic space with their reference representation β^r. E.g., in Figure 1, two topics $\mathcal{C}_1$:*sports* and $\mathcal{C}_2$:*politics* represents two axes (i.e., reference axes) defined by their reference representations β_1^r and β_2^r respectively. Defining a target corpus by these two topics is analogous to finding a representation in the space defined by the reference axes. Figure 1 shows such target representation β^{USA} and β^{INDIA} as coordinates by reference axes for corpora specific to USA and INDIA, respectively.

In practice, the above interpretation of defining a corpus can be helpful from several perspectives. First, it has better **interpretability** than traditional document modeling approaches like topic models (Blei et al., 2003; Jordan et al., 1999) that discover some unknown topics. It defines a corpus over well-defined topics by finding their corpus-specific representations. Thus, users do not need much ef-

[1]Code and data are available at https://github.com/pritomsaha/Coordinated-Topic-Modeling

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9831 - 9843
December 7-11, 2022

fort to understand unknown topics; instead, they easily grasp the corpus-specific aspects of given well-known topics. Second, it facilitates the **comparability** between two corpora. It uses a predefined set of topics with their reference representation as the axes to describe a corpus in a semantic space. Thus, two corpora represented by the same axes can be easily comparable by studying their relative position in the coordinates. Third, it allows **reusability** by utilizing existing information to explore new knowledge. As the reference representation is already known information about some topics, resuing it to model a new corpus in a similar domain is intuitive and helpful. Specifically, this is useful for small target corpora as traditional models are not typically effective in such cases.

Considering the above motivations, we thus formalize a new problem, **Coordinated Topic Modeling** (CTM). It takes a target corpus $\mathcal{D}$, a reference representation β^r for k well-defined topics with their surface names $\mathcal{C}$. As output, it aims to learn a target representation β to best define $\mathcal{D}$. In many cases, we can obtain a set of well-defined topics with their representation. E.g., there may exist some large corpora in a particular domain annotated with topics, and there are effective existing methods (Ramage et al., 2009, 2011) to obtain the reference representations. Specifically, we use labeled LDA (Ramage et al., 2009) to get reference representation (more details on Section 3.1)

Some previous attempts incorporate prior knowledge into topic models to impose more interpretability. One such work takes document-level supervision by providing topic labels of all or a subset of documents (Mcauliffe and Blei, 2007; Ramage et al., 2009). While they improve the predictive ability of unsupervised topic models, they require a massive set of annotated documents to be effective. An alternative way is called topic-level supervision by providing seed words for each topic (Eshima et al., 2020; Harandizadeh et al., 2022). Although they make the topics converge toward the user's interest, the seed words should be only from the target corpus vocabulary, which may be impractical in many cases. Finally, the category-guided topic mining (Meng et al., 2020a) considers the topic names as the only supervision for mining user-desired discriminative topics. However, it also assumes that the name of each topic appears in the corpus. Moreover, none of the above works does impose requirements of making topics comparable over multiple corpora (see Appendix A for a detailed discussion on related work).

From the above discussion, we identify the following two **requirements** that the solution of CTM needs to meet. **(1)** The target representation should be learned based on reference representation by capturing the target corpus-specific aspects; **(2)** It also needs to relate the documents in the corpus with the global semantics of each topic represented by its surface name so that it maintains general interpretability and comparability. Although previous work considers mining topics from user guidance, none of those fulfill these two requirements.

There are several challenges to fulfilling the above requirements. The **first challenge** is *handling the vocabulary mismatch* problem. As the vocabulary of β^r and $\mathcal{D}$ can be the different, β^r cannot be directly calibrated into the target β. To solve this problem, we propose a method called *reference projection*. Here, the main idea is first to generate a proxy representation $\tilde{\beta}^r$ of the target β having the same vocabulary with reference β^r, making it comparable for supervision. It then enforces $\tilde{\beta}^r$ and β^r to be as close as possible, which indirectly allows the target β to be guided by given β^r even if they have different vocabularies. Moreover, another **benefit** of $\tilde{\beta}^r$ is that it enables directly comparing two corpora given the same topics set.

The **second challenge** is *providing surface names guidance*. We only know the surface name of each given topic without further knowledge of how each document corresponds to them. Therefore, we generate document-level supervision from the surface names using the *textual entailment* approach (Yin et al., 2019). It imitates how humans determine the topic(s) of a document by filling in a template (e.g., "this document is about <topic name>"). In this paper, we utilize a pre-trained textual entailment model (Liu et al., 2019) with given surface names to generate document-level distribution matrix θ^t. This θ^t later guides CTM by relating the global semantics of surface names with given documents.

The **final challenge** is *combining two supervisions*. Now, we have topic-level supervision from prior projection and document-level supervision from the textual entailment model. To combine these two supervisions, we **propose** a method called *embedding based coordinated topic model*, **ECTM**. It exploits the architecture of an embedded topic model (ETM) (Dieng et al., 2020). The main

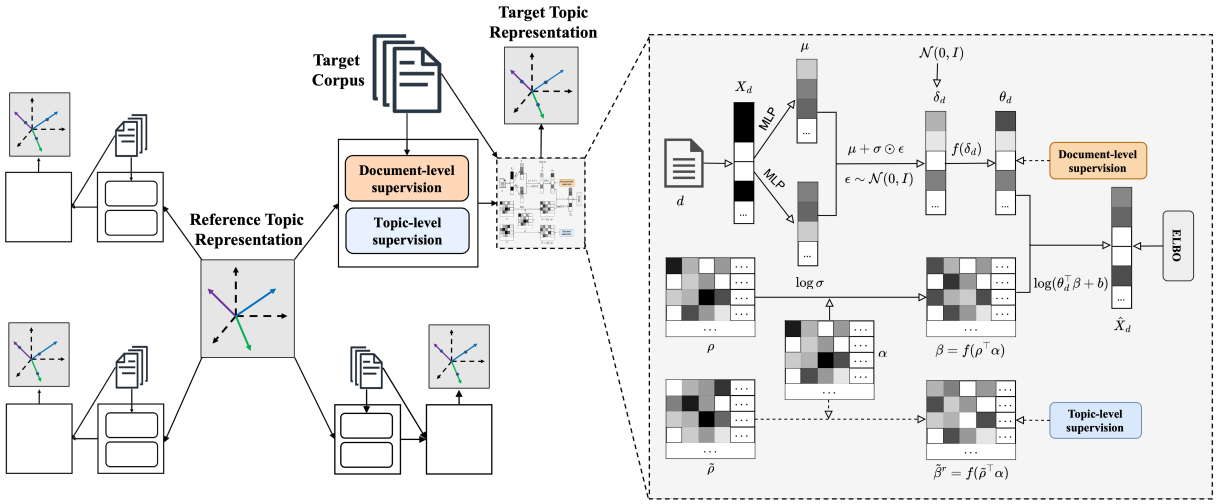

Figure 2: Proposed Architecture

idea is to regularize the objective of ETM using our two proposed supervisions. To further generalize ECTM, we employ a mechanism similar to *self-training* (Meng et al., 2020b) where we iteratively use the model's current output to update θ^t.

Our **contributions** can be summarized as follows. **First**, we propose *coordinated topic modeling*, a new problem for modeling a text corpus using well-defined topics as reference. **Second**, we develop an embedding-based framework for solving the problem. It uses given reference representation to effectively model a corpus while maintaining the semantics of given surface names. **Third**, We propose methods for generating topic-level and document-level supervisions using an introduced projection mechanism and the textual entailment approach, respectively. We then combine these two supervisions into a unified model to solve the CTM problem. **Finally**, we conduct a comprehensive set of experiments on multiple domains for different tasks, demonstrating our framework's superiority against strongly designed baselines.

2 Proposed Methodology

Overview. To solve the CTM problem, we design a method ECTM (Sec. 2.2) based on an embedded topic model (ETM) (Dieng et al., 2020) by imposing our requirements. First, we incorporate a topic-level supervision (Sec. 2.2.1) from given reference representation β^r by introducing a mechanism called *reference projection*. It generates a proxy representation of intended β with the same vocabulary dimension with β^r, thus enabling supervision. Second, we include a document-level supervision (Sec. 2.2.2) from the given surface name of each topic by employing the *textual entailment* approach (Yin et al., 2019). Finally, we combine two supervisions (Sec. 2.2.3) by regularizing the ETM's objective, To further generalize the model, we employ a self-training (Meng et al., 2020b) by iteratively using the model's current output to update the supervision. The rest of the section first reviews ETM and then presents our ECTM.

2.1 Embedded Topic Model

ETM is a neural topic model that uses vector representation of both words and topics to improve topic quality and predictive accuracy of LDA (Blei et al., 2003). Let $\rho \in \mathbb{R}^{L \times V}$ and $\alpha \in \mathbb{R}^{L \times K}$ be L-dimensional embeddings of V vocabulary words and K latent topics respectively, ETM defines k^{th} topic $\beta_k = f(\rho^\top \alpha_k)$ where $f(\cdot)$ is the softmax function. It uses α in its generative process of d^{th} document of corpus $\mathcal{D}$ as follows:

1. Draw topic proportion $\theta_d \sim \mathcal{LM}(0, I)$ where $\mathcal{LM}(\cdot)$ is a logistic normal distribution (Atchison and Shen, 1980) that transforms a standard Gaussian random variable to the simplex. In other word, $\theta_d = f(\delta_d)$ where $\delta_d = \mathcal{N}(0, I)$.
2. For each word n in the document:
   (a) Draw topic assignment $z_{dn} \sim Cat(\theta_d)$.
   (b) Draw the word $w_{dn} \sim f(\rho^\top \alpha_{z_{dn}})$.

Here, $Cat(\cdot)$ denotes the categorical distribution. ETM employs variation inference (Jordan et al., 1999; Blei et al., 2017) that uses the evidence lower bound (ELBO) as a function of the model parameters (α, ρ) and the variational parameters v:

$$\mathcal{L}(\Theta) = \sum_{d\in\mathcal{D}}\sum_{n=1}^{N_d}\mathbb{E}_q[\log p(w_{dn} \mid \delta_d, \rho, \alpha)] - \sum_{d\in\mathcal{D}} KL(q(\delta_d; w_d, v) \mid\mid p(\delta_d)), \quad (1)$$

where Θ represents the model and variational parameters. $q(\cdot)$ is a Gaussian whose mean, and variance are estimated from a neural network.

As a function of model parameters, ELBO in Eq. (1) is equivalent to the expected complete log-likelihood maximization. On the other hand, the first term in Eq. (1) encourages variational parameters to place mass on topic proportions δ_d that explains the observed words, and the second term forces them to be as close as possible to $p(\delta_d)$.

2.2 ECTM

The proposed ECTM uses ETM as the base model to solve the CTM problem. There are several reasons for this. Firstly, ETM combines the strength of the neural topic model and word embedding for modeling corpus more effectively. Secondly, it lets us use pre-trained word embedding to map words in a common vector space even if the words do not appear in the target corpus vocabulary. Thirdly, we can regularize the objectives of ETM by imposing our problem-specific requirements.

Figure 2 shows the base model, excluding the parts involving dashed lines. Similar to ETM, we use a two-layer perceptron to encode the bag of words (BOW) representation X_d of a document d into document-topic distribution θ_d. At the same time, vector representation of words ρ and topics α generate topic-word distribution β. Finally, from θ_d and β, X_d is reconstructed as $\hat{X}_d = \log(\theta_d^\top \beta + b)$. where b is the *background bias* representing the relative frequency of the vocabulary words in $\mathcal{D}$. Unlike ETM, we include b to account for words with approximately the same frequency across documents. It helps produce coherent topics by weighing down common words. Based upon this base model, ECTM consists of the following three components for incorporating the CTM's requirements.

2.2.1 Topic-level Supervision

Our first requirement stems from the fact that, in many cases, describing a corpus with some arbitrary topics incurs a burden for users to understand them first. Instead, it is more convenient to summarize the corpus over some well-defined topics. Thus, in CTM, we consider a set of well-defined topics with reference representation β^r (i.e., word distribution) like the axes in semantic space by

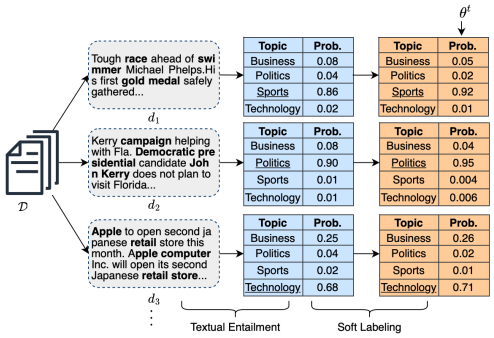

Figure 3: Document-level Supervision Generation

employing supervision to generate a target representation β that best describes the given corpus $\mathcal{D}$. Here, β^r may come from existing sources. E.g., a large corpus in a similar domain annotated with topics can be used by supervised topic models as they are effective for finding high-quality topics from a large annotated corpus.

Now, the question is, how can we use β^r as reference axes in the semantic space to guide generating β? One possible solution can be constraining the ETM so that the generated β comes close to β^r while also maximizing ELBO. However, the problem is that we cannot assume that β^r and β share the same vocabulary. Hence, β^r cannot be used directly as guidance. To solve this problem, we take an indirect way of providing supervision which we call *reference projection*.

Reference Projection. As shown in Figure 2, alongside β, we also generate a representation $\tilde{\beta}^r = f(\tilde{\rho}^\top \alpha)$. Here, $\tilde{\rho}$ is the embedding matrix of the vocabulary that the given reference β^r is based on. Now, we aim to enforce $\tilde{\beta}^r$ and β^r to be as close as possible by minimizing the following:

$$R_\beta = \frac{1}{k}\sum_{j=1}^{k} KL(\beta_j^r, \tilde{\beta}_j^r). \quad (2)$$

where k is the number of topics. Here, minimizing R_β makes the model update the topic embedding matrix α, which involves generating β. Thus, it indirectly guides β even if there is a vocabulary mismatch between β and β^r. In other words, β is a kind of projection of β^r on the target corpus vocabulary dimension.

2.2.2 Document-level Supervision

According to the second requirement, we want to maintain a global semantic of the given topic represented by its surface name. The reason is that if the target representation of each topic deviates much from its well-defined meaning, it may compromise the interpretability.

For this, we first try to understand the semantics of a document by relating it with given surface names. Mainly, we use a textual entailment approach (Yin et al., 2019) to calculate the probability of a document belonging to a topic (i.e., surface name). As humans, based on a document's content, we can determine the topic(s) by asking to what extent that document belongs to a topic out of some options. Similarly, using the textual entailment approach, we imitate humans to determine the probability that a document is about a topic by creating a hypothesis by filling in the surface name in a template (e.g., "this document is about <surface name>"), given the context document.

Given the surface names $\mathcal{C}$, we generate θ^t for document-level supervision by using a pre-trained textual entailment model (Liu et al., 2019) fine-tuned on Yahoo Answers topic classification[2]. It considers an input document d as the "premise", creates a "hypothesis" made of a template filled with a $c_k \in \mathcal{C}$, and generates a probability p_{dk} denoting to what extent the premise entails the hypothesis. We have two major choices of θ^t from these generated probabilities. Firstly, we can use hard labeling (Lee et al., 2013) which converts high-probability over a threshold τ to one-hot labels, i.e., $\theta^t_{dk} = 1(p_{dk} > \tau)$ where $1(\cdot)$ is the indicator function. Secondly, we can use the generated p_{dk} as it is for θ^t_{dk} as a soft label. In soft labeling, we can further use an approach (Bhatia et al., 2016) which elevates the high-probability label while demoting the low-probability ones. More specifically, it squares and normalizes the p_{dk} as follows:

$$\theta^t_{dk} = \frac{p^2_{dk}/f_k}{\sum_{k'} p^2_{dk'}/f_{k'}}, \quad f_k = \sum_{d \in \mathcal{D}} p_{dk}. \tag{3}$$

We observe that the soft labeling strategy consistently performs better and provides more stable results than the hard labeling. The likely reason is that as hard labeling considers high-probability topics as direct labels, it is more susceptible to error propagation. Moreover, another drawback of hard labeling is that we need to set a threshold explicitly, while soft labeling does not require that. Figure 3 shows the overview of generating θ^t.

The θ^t is then used to provide document-level supervision by minimizing the following :

$$R_\theta = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} KL(\theta^t_d, \theta_d). \tag{4}$$

[2] https://huggingface.co/joeddav/bart-large-mnli-yahoo-answers

Algorithm 1: ECTM

Input: A target corpus $\mathcal{D}$; a set of topics with surface names $\mathcal{C}$ and reference representation β^r; a pre-trained word embedding $\mathcal{W}$; an entailment model $\mathcal{E}$.
Output: Trained Model $\mathcal{M}$ with β.

```
Use E to initialize θ^t ← Section 2.2.2;
From W, obtain ρ and ρ̃;
B ← Total number of batches;
for i ← 0 to B − 1 do
    Train model M on D with Eq. (5);
    if i mod 50 = 0 then
        Update θ^t with Eq. (6);
    end if
end for
return M;
```

As we use an existing knowledge source pre-trained on a massive volume of documents over various domains, this supervision helps maintain the global semantics of each topic in its target representation. Moreover, it also improves the model's predictive power to identify its topic.

2.2.3 Unification and Self-training

Now, we have topic-level and document-level supervision R_β and R_θ respectively by imposing CTM's requirements. We unify them into one model by constraining the objective of our base model as follows:

$$\mathcal{L}(\Theta) = ELBO - \lambda_\beta R_\beta - \lambda_\theta R_\theta, \tag{5}$$

where λ_β and λ_θ are the regularization weights for R_β and R_θ respectively. Maximizing Eq. (5) jointly ensures the following objectives: (1) The ELBO part enforces the model to explain $\mathcal{D}$ by reducing the reconstruction error; (2) R_β encourages the model to converge β in the direction of β^r; and (3) R_θ encourages the model to maintain the global semantics of given coordinates in β by enforcing θ and θ^t as close as possible.

Finally, to further generalize the model's predictive strength in θ, we use a *self-training* (Meng et al., 2020b) like mechanism. Here, the main idea is to iteratively use the model's current θ to update θ^t for supervision. More specifically, we update θ^t after every 50 iteration as follows:

$$\theta^t = 0.5 * \theta^t + 0.5 * \theta. \tag{6}$$

It also reduces the chance of error propagation from the textual entailment model. We summarize our ECTM framework in Algorithm 1.

3 Experiments

In this section, we employ empirical evaluations, which are designed mainly to answer the following research questions (RQs): **RQ1.** How effective is the ECTM quantitatively in terms of the quality of topics for the CTM problem and text classification performance? **RQ2.** How does the ECTM qualitatively perform in terms of interpretability and distinctiveness of generated topic words? **RQ3.** How does ECTM benefit the task of comparing multiple corpora? **RQ4.** How does each part of ECTM contribute to its performance?

3.1 Experiment Setup

Datasets. We use datasets from three different domains *news articles*, *review sentiment* and *academic articles* domains. For news articles, we obtain three datasets: (1) 20 Newsgroup corpus[3] (**20Newsg**); (2) New York Times annotated corpus (**NYT**) (Sandhaus, 2008); (3) AG's News dataset (**AGNews**) from (Yang et al., 2016). For review sentiment domain, we use: (1) Yelp restaurant review dataset used in (Meng et al., 2020a) (**Yelp-Sent**); (2) IMDB Movie Review dataset used in (Hoang et al., 2019) (**IMDB-Sent**). Finally for academic articles: (1) Arxiv Artificial Intelligence (AI) article abstracts spanning 2020-2022 [4] (**Arxiv-AI**); (2) Microsoft Academic Graph AI article abstracts (Sinha et al., 2015) (**MAG-AI**).

Baselines. We compare our model with the following baselines.

- GLDA: Guided LDA (Jagarlamudi et al., 2012) biases LDA's generative process using topic-level priors over vocabulary from given seed words.
- Sup+LLDA: Supervised Labeled LDA is based on Labeled-LDA (Ramage et al., 2009) where a label for each document is predicted from a supervised BERT classifier [5] learned on annotated reference corpus.
- ZS+LLDA: Zero-Shot Labeled LDA is also based on Labeled-LDA (Ramage et al., 2009) where a label for each document is inferred from given surface names using a Zero-Shot classification (Yin et al., 2019).
- ACorEx: Unlike LDA, Anchored CorEx (Gallagher et al., 2017) is not based on generative assumptions and uses topic correlation to learn maximally informative topics. Moreover, it uses user-provided seed words as anchors to bias compression of the original corpus.
- AVIAD: AVIAD (Hoang et al., 2019) extends the Autoencoding Variational Inference for Topic Models (AVITM) (Srivastava and Sutton, 2017) approach to incorporate prior knowledge from seed words by modifying the loss function to infer desired topics. It uses a variational autoencoder like ETM but does not employ word embedding in the modeling.
- KeyETM: Keyword Assisted ETM (Harandizadeh et al., 2022) modifies ETM's objective to include prior knowledge from given seed words.

Methods	20Newsg			NYT			Yelp-Senti			Arxiv-AI		
	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ
GLDA	0.25	0.87	0.22	0.26	0.85	0.22	0.08	0.80	0.06	0.09	0.93	0.09
Sup+LLDA	0.23	0.79	0.18	0.20	0.63	0.12	0.06	0.70	0.04	0.04	0.46	0.02
ZS+LLDA	0.23	0.80	0.18	0.17	0.65	0.11	0.06	0.76	0.05	0.14	0.80	0.11
ACorEX	0.25	**1.00**	0.25	0.27	**1.00**	0.27	0.07	**1.00**	0.07	-0.03	0.96	-0.03
AVIAD	0.13	**1.00**	0.13	-0.26	**1.00**	-0.26	-0.01	**1.00**	-0.01	-0.34	**1.00**	-0.34
KeyETM	0.26	**1.00**	0.26	0.19	0.89	0.17	0.07	0.92	0.07	0.04	**1.00**	0.04
ECTM	**0.30**	**1.00**	**0.30**	**0.28**	0.97	**0.27**	**0.09**	**1.00**	**0.09**	**0.15**	0.97	**0.15**

Table 1: Quality Measures of Topic

Reference Topic Representation. As an input, CTM requires prior knowledge β^r. For each domain, we use a large annotated corpus as a reference to get β^r using labeled LDA (LLDA) (Ramage et al., 2009). Because LLDA is effective in producing high-quality topics when there is a large annotated corpus. E.g., in news domain, we use a large annotated corpus **AGNews** as reference with topic names *business*, *politics*, *sports* and *technology*. In the sentiment domain, the reference corpus we used is **IMDB** with positive (*good*) and negative (*bad*) sentiments. Finally, we use the **MAG-AI** as reference corpus for the AI domain to generate prior, annotated with four topics: *computer vision* (CV), *information retrieval* (IR), *machine learning* (ML), and *natural language processing* (NLP).

Seed words. For the baselines that require seed words, we collect them from given surface names $\mathcal{C}$ and top topic words based on given β^r. If a word does not appear in the corresponding target corpus, we replace that with the most similar word using cosine similarity. As suggested in the baselines, we provide 10 seed words for each topic.

The implementation details of our model and baselines are specified in Appendix B.

3.2 Topic Quality

Evaluation Metrics. We use the following three well-defined quantitative measurements to evaluate

[3] http://qwone.com/~jason/20Newsgroups/

[4] https://www.kaggle.com/Cornell-University/arxiv

[5] https://huggingface.co/docs/transformers/tasks/sequence_classification

	20Newsg		**NYT**		**Yelp-Senti**		**Arxiv-AI**	
	sports	politics	business	technology	good	bad	ML	IR
Reference Topic Words	night	leader	stock	software	song	waste	machine	retrieval
	play	election	sale	technology	music	awful	learning	document
	sport	attack	share	service	musical	terrible	algorithm	query
	player	afp	billion	internet	wonderful	boring	optimization	search
	beat	iraqi	fall	launch	dance	poor	problem	base
GLDA	game	people	company	president (×)	good	order (×)	adversarial	graph
	team	time (×)	percent	bush (×)	place (×)	food (×)	distribution (×)	search
	year (×)	government	year (×)	official (×)	food (×)	time (×)	class (×)	user
	play	gun	bank	united (×)	great	place (×)	function (×)	class (×)
	player	year (×)	market	house (×)	order (×)	service (×)	attack (×)	recommendation
Sup+LLDA	game	people	year (×)	year (×)	place (×)	food (×)	demonstrate (×)	retrieval
	team	government	percent	time (×)	food (×)	order (×)	problem (×)	exist (×)
	year (×)	kill	company	people (×)	good	place (×)	feature	demonstrate (×)
	play	time (×)	market	president (×)	great	service (×)	training	representation
	time (×)	year (×)	government (×)	official (×)	service (×)	time (×)	neural	feature (×)
ZS+LLDA	game	people	year (×)	year (×)	food (×)	food (×)	efficient (×)	retrieval
	team	time (×)	percent	time (×)	place (×)	order (×)	reduce (×)	search
	year (×)	government	company	american (×)	good	place (×)	number (×)	user
	play	year (×)	market	official (×)	great	service (×)	leverage (×)	document
	player	point (×)	lead (×)	today (×)	service (×)	time (×)	module (×)	query
ACorEX	point	force	billion	release (×)	good	bad	optimization	search
	play	country	business	technology	hear (×)	money (×)	gradient	document
	player	attack	buy	phone	beautiful	terrible	convergence	query
	league	military	stock	time (×)	music (×)	poor	stochastic	retrieval
	beat	political	profit	space	sound (×)	waste	print (×)	semantics
AVIAD	robitaille (×)	tragedy (×)	sanwa (×)	genscher (×)	traditional (×)	email (×)	bind	ehr
	probert (×)	policy	zoete (×)	enlargement (×)	snow (×)	upset	analytically (×)	healthy (×)
	howe (×)	serbian	earning	abm (×)	filling	management (×)	certify (×)	progression (×)
	player	freedom	overprice	teng (×)	bisque (×)	yell	arm (×)	patient (×)
	nhl	unite (×)	acquirer	chechnya (×)	seaweed (×)	acknowledge (×)	pruning	ehrs
KeyETM	game	people	year (×)	company (×)	good	food (×)	function	translation (×)
	team	government	percent	bank (×)	place (×)	order (×)	estimation (×)	user
	season	person (×)	market	japan (×)	great	service (×)	distribution (×)	search
	play	armenian	time (×)	china (×)	time (×)	eat (×)	parameter (×)	annotation
	win	law	month (×)	russia (×)	love	restaurant (×)	efficient (×)	point (×)
ECTM	game	government	company	space	great	waste	optimization	retrieval
	team	war	bank	site	music (×)	awful	convergence	document
	win	military	percent	technology	love	terrible	stochastic	query
	season	armenian	market	station	wonderful	bad	gradient	search
	league	attack	price	network	amazing	horrible	function	user

Table 2: Qualitative Evaluation

the inferred topics' quality:

- Topic coherence (TC) is a standard measure of interpretability based on the average point-wise mutual information between randomly drawn two words from the same document (Lau et al., 2014).
- Topic diversity (TD) measures the percentage of unique words in the top 25 words of all topics (Dieng et al., 2020). It captures the semantical diverseness of the inferred topics.
- Topic quality (TQ) (Dieng et al., 2020) is the overall metric for measuring the quality of topics as the product of topic coherence and diversity.

Results and Discussions. We first show the quantitative results of topic quality in Table 1. The results suggest that, in general, ECTM generates more coherent and interpretable topics than other baselines. In some cases, other methods produce more diverse topics than ours. E.g., ECTM's generated topics' diversity scores are slightly lower than the best one in **NYT** and **Arxiv-AI** datasets. However, our method significantly outperforms others in quality scores in those cases by producing more coherent topics. Thus, ECTM produces more interpretable topics while also maintaining diversity.

In Table 2, we show randomly selected two topics from each dataset and show the top-5 words under each topic. We also show the top-5 words from the given β^r. Words that authors determined to be irrelevant to the corresponding topic are marked with (×). Overall, our method-generated topic words are relevant and easily interpretable in nearly all cases compared to baselines. We also observe that the topics produced by AcorEx are reasonably good in terms of interpretability. However, AcorEx's produced topics strictly converge toward the prior representation rather than adapting according to the target corpus. E.g., in *sports* topic, 3 out of 5 topic words overlap with priors topic words. In contrast, our method tends to capture the target corpus-specific aspects of the given topics.

On the other hand, AVIAD suffers from the opposite issue. It adapts too much that the topics become very difficult to understand. The KeyETM with a similar base model (ETM) to ours performs better when the target corpus is balanced. E.g., as **20Newsg** is a comparatively balanced dataset, the keyETM performs well. In contrast, our proposed ECTM consistently performs better as it enjoys the benefit of both topic-level supervision from β^r and document-level supervision from existing knowledge sources to make the topics adjusted to the target corpus as well as not going away from

Methods	20Newsg			NYT			Yelp-Senti			Arxiv-AI		
	Acc.	F1-Mac	F1-Mic	Acc.	F1-Mac	F1-Mic	Acc.	F1-Mac	F1-Mic	Acc.	F1-Mac	F1-Mic
GLDA	0.70	0.71	0.73	0.70	0.59	0.75	0.75	0.75	0.75	0.67	0.59	0.71
Sup+LLDA	0.68	0.53	0.62	0.77	0.70	0.80	**0.94**	**0.94**	**0.94**	0.38	0.37	0.38
ZS+LLDA	0.89	0.80	0.88	0.91	0.79	0.90	0.87	0.87	0.87	0.34	0.34	0.33
ACorEX	0.38	0.43	0.44	0.72	0.63	0.73	0.60	0.59	0.58	0.66	0.51	0.66
AVIAD	0.72	0.71	0.74	0.69	0.52	0.72	0.75	0.74	0.75	0.64	0.57	0.68
KeyETM	0.66	0.62	0.69	0.39	0.31	0.40	0.49	0.34	0.32	0.30	0.24	0.23
SupBERTCLs	0.68	0.53	0.62	0.77	0.70	0.80	**0.94**	**0.94**	**0.94**	0.38	0.37	0.38
Zero-ShotCLs	0.89	0.80	0.88	0.91	0.79	0.90	0.87	0.87	0.87	0.34	0.34	0.33
WesSTCLs (Names)	0.55	0.47	0.57	0.72	0.65	0.76	0.80	0.79	0.79	0.44	0.44	0.48
WesSTCLs (Seeds)	0.73	0.65	0.76	0.79	0.70	0.83	0.64	0.64	0.64	0.79	0.68	0.81
LOTCLs	0.84	0.68	0.81	0.84	0.75	0.86	0.76	0.75	0.75	0.07	0.09	0.06
ECTM	**0.91**	**0.85**	**0.91**	**0.91**	**0.83**	**0.92**	0.90	0.90	0.90	**0.83**	**0.76**	**0.83**

Table 3: Text Classification Performance

well-known semantics of given topics' names.

3.3 Text Classification

Although the primary purpose of our model is not document categorization, we can use learned θ to analyze the document representation. More specifically, we can consider the topic with maximum probability θ_d as the document's label of d. Therefore, we can evaluate how learned topics are distinctive and informative enough to represent a document to categorize it correctly. Alongside the topic models, to compare text classification performance, we also include several supervised and semi-supervised baselines: SupBERTCLs, Zero-ShotCLs (Yin et al., 2019), WeSTCLs (Meng et al., 2018), and LOTCls (Meng et al., 2020b), with details described in Appendix C.

Evaluation Metrics. We use accuracy (Acc.), macro-F1 (F1-Mac) and micro-F1 (F1-Mic) scores that are commonly used in classification evaluations (Meng et al., 2018, 2020b) as the metrics.

Results and Discussions. The text classification results on **20Newsg**, **NYT**, **Yelp-Senti** and **Arxiv-AI** datasets are shown in Table 3. We can see that ECTM mostly performs better by all three evaluation metrics than other models with large margins. One reason for this better result is the document-level supervision we employ from the textual entailment model. It enforces the model to generate θ with better predictive power. It can also be explained by the comparable performance with the baselines ZS+LLDA and Zero-ShotCLs models, as they also use the textual entailment approach. Here, we can see Sup+LLDA and SupBERTCLs perform slightly better than ours in the Yelp-Sent dataset because it is rather an easy task to learn a classifier of binary sentiment classes from supervision. However, it performs very poorly for other more hard classification datasets.

Moreover, our model performs better even when the textual entailment model fails to perform well. E.g., in **Arxiv-AI** dataset, while the ZS+LLDA and Zero-ShotCLs have less than 0.40 for all three

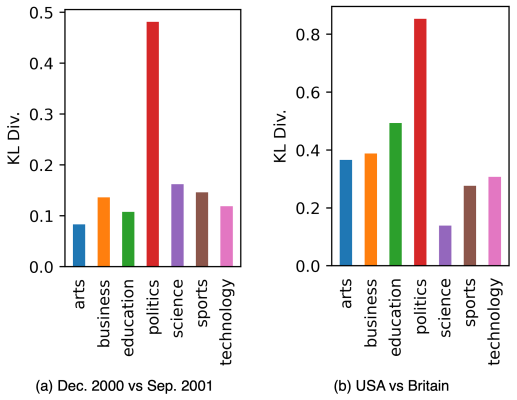

Figure 4: Corpus Comparison

Dec. 2000	court, election, vote, bush, president, florida, judge, supreme
Sep. 2001	attack, war, president, military terrorism, terrorist, afghanistan, muslim
USA	republican, president, governor, senator, clinton, bush, giuliani, senate
Britain	prime, minister, ireland, irish, northern, blair, union, thatcher

Table 4: Context defining words in politics topic

measures, ECTM performs significantly better by attaining close or more than 0.80 scores in the evaluation measures. This is because our model does not rely solely on the textual entailment model. It self-trains itself by updating supervision iteratively by reducing the chance of error propagation from the entailment model.

3.4 Corpus Comparison

This section explores how ECTM facilitates the quantitative comparison of any two corpora given the same topics. In ECTM, as described before, while incorporating topic-level supervision, we generate a proxy topic-word distribution $\tilde{\beta}^r$ of the target β to make it comparable with β^r. As both β^r and $\tilde{\beta}^r$ share the same vocabulary, any two corpora with the same sets of topics can be thus quantitatively compared using their generated $\tilde{\beta}^r$s. To investigate the comparison, we apply our ECTM to multiple corpora over different contexts using the same topics. More specifically, we use two contexts: time and location. For time context, we use NYT articles from December 2000 and September 2001 as two target corpora. For location context, we use NYT articles on USA and Britain as two target corpora. In both cases, we use the same set of topics, and the entire NYT corpus is used to obtain the reference.

Let $\tilde{\beta}^{r(1)}$ and $\tilde{\beta}^{r(2)}$ are generated by two corpora $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, we compare $\mathcal{D}_1$ and $\mathcal{D}_2$ by calculating $KL(\tilde{\beta}_j^{r(1)} || \tilde{\beta}_j^{r(2)})$ for each topic $j \in \{1, \cdots, k\}$. Figure 4 illustrates such results for our two contexts. Here, we see in both con-

Methods	Quality			Classification			Top 5 Words of Topic
	TC	TD	TQ	Acc.	F1	F1	Business
ECTM - R_β - R_θ - b	0.24	0.99	0.24	–	–	–	time, work, year, problem, people
ECTM - b	0.28	0.98	0.27	0.90	0.83	0.90	year, president, **price**, month, week
ECTM - R_β	0.20	0.99	0.19	**0.92**	**0.87**	**0.92**	fan, **sale**, la, great, **pay**
ECTM - R_θ	0.30	0.99	0.30	0.83	0.71	0.82	president, **price**, **buy**, **stock**, **oil**
ECTM	**0.30**	**1.0**	**0.30**	0.91	0.85	0.91	**price**, **stock**, **sale**, **buy**, **oil**

Table 5: Ablation Study

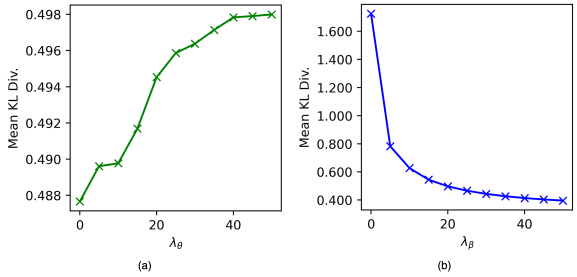

Figure 5: Effect of (a) Document-level supervision and (b) Topic-level supervision on topic's adaptability to the target corpus

texts, the topic *politics* has made the highest difference between the two corresponding corpora. We also show some context defining words out of top 20-words obtained from ECTM generated β for politics in Table 4 (full list is in Appendix D). As December 2000 was the month when the *Bush vs. Gore 2000 presidential race* was settled, the words like *election*, *vote* or *judge* make sense. And words like *attack*, *terrorism* or *war* are dominated in *politics* topic on September 2001 as there was a terror attack that month. Similarly, the words that mainly make the difference in politics for USA and Britain are self-explanatory.

3.5 Ablation Study

In this section, we investigate the role of different parts of our proposed model in **20Newsg** dataset (shown in Table 5). First, when we exclude the background bias b from the model, we see the common words (i.e., *time*, *year*) are being dominated in the top-5 topic words, thus making it less interpretable and distinctive. It explains the effectiveness of background bias in the model. Second, when only the document-level supervision (ECTM - R_β) is employed, the model's predictive power gets improved. However, the topic quality scores are significantly downgraded, which is also reflected in irrelevant words (i.e., general words like *great, la*) in the top-5 topic words. It justifies the intuition of our topic-level supervision to make topics interpretable by calibrating them from well-defined topics. On the other hand, only using topic-level supervision degrades the model's predictive power. Finally, the complete model balances them by generating interpretable and discriminatory topics, which is reflected in the quantitative measures.

Moreover, in Figure 5, we show the effect of two supervisions on the topic's adaptability in the target corpus. The vertical axis shows the divergence of learned topics with the given reference topics. If the divergence score increases, it means topics are getting more adapted to the corpus than being similar to the reference one. Figure 5 (a) shows that if we increase the regularization weight parameter λ_θ for document-level supervision by fixing λ_β, the topics get deviated from reference topic focusing more on the target corpus content. On the other hand, while increasing topic-level regularization weight λ_β by fixing λ_θ, the model tends to converge towards the reference topics. It is intuitive and reasonable according to our discussion of model requirements.

4 Conclusion

In this paper, we propose a new problem called coordinated topic modeling that uses a set of well-defined topics to describe a target corpus. It takes a reference representation of topics with their surface names and calibrates those representing the target corpus. The proposed problem describes corpora in a more interpretable and comparable way. To solve the problem, we design an embedding-based coordinated topic model that leverages topic-level and document-level supervision with a self-training mechanism. A set of empirical evaluations demonstrate the superiority of our approach over other strong baselines in multiple datasets.

Acknowledgements

We thank the reviewers for their constructive feedback. This material is based upon work supported by the National Science Foundation IIS 16-19302 and IIS 16-33755, Zhejiang University ZJU Research 083650, IBM-Illinois Center for Cognitive Computing Systems Research (C3SR)– a research collaboration as part of the IBM Cognitive Horizon Network, grants from eBay and Microsoft Azure, UIUC OVCR CCIL Planning Grant 434S34, UIUC CSBS Small Grant 434C8U, and UIUC New Frontiers Initiative. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Limitations

The proposed model can be applied only to those domains where we can find a reference representation of intended topics. However, we aim for the topic's interpretability and comparability with other corpora. Thus, in this work, we are mainly interested in defining a corpus with well-defined topics as input because they are easily understandable to represent a corpus. Moreover, it is very reasonable to think that we can usually obtain reference representation for well-defined topics.

In this paper, the model assumes that the reference and target corpus are from a common domain. However, it is worth exploring how our model works in cross-domain problems. On the other hand, exploring the proposed model in cross-domain scenarios is also an excellent future direction. We can apply our model in cross-language applications if word embeddings in the vocabulary are available for multiple corpora from different languages.

References

J Atchison and Sheng M Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

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

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.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2020. Keyword assisted topic models. *arXiv preprint arXiv:2004.05964*.

Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16.

Bahareh Harandizadeh, J. Hunter Priniski, and Fred Morstatter. 2022. Keyword assisted embedded topic model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 372–380, New York, NY, USA. Association for Computing Machinery.

Tai Hoang, Huy Le, and Tho Quan. 2019. Towards autoencoding variational inference for aspect-based opinion summary. *Applied Artificial Intelligence*, 33(9):796–816.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Simon Lacoste-Julien, Fei Sha, and Michael Jordan. 2008. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in neural information processing systems*, 21.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020a. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pages 2121–2132.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992. ACM.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020b. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923.

A Related Work

It is of great interest to automatically mine a set of meaningful and coherent topics for effectively and efficiently understanding and navigating a large text corpus. Topic models (Jordan et al., 1999; Blei et al., 2003) are such statistical tools that can discover latent semantic themes from a text collection. The main idea is to represent each document in a corpus as a mixture of hidden topics, each represented by word distribution. While most of the early attempts of topic models (Griffiths et al., 2003; Blei and Lafferty, 2006; Li and McCallum, 2006) are probabilistic, with the advent of the deep neural network, neural topic models are also proposed. One such example is the Autoencoded Variational Inference For Topic Model (AVITM) (Srivastava and Sutton, 2017). Recently, Embedded Topic Model (ETM) (Dieng et al., 2020) blends the strengths of the neural topic model and word embedding.

Despite their effectiveness and efficiency, traditional topic models have several limitations. The standard topic models whether it is probabilistic (Blei et al., 2003) or neural (Srivastava and Sutton, 2017; Dieng et al., 2020) has the inability to incorporate user guidance. Existing topic models are typically learned in a purely unsupervised manner and, thus, tend to discover the most general and major topics ignoring user interests. There have been several modifications in the traditional topic models to incorporate user interests or existing knowledge about documents in the corpus. More specifically, there are generally three forms of supervision in existing supervised topic models.

The first type of supervision is in the form of labeled corpus where all or a subset of documents are annotated with some predefined topic, or class labels (Ramage et al., 2009, 2011; Mcauliffe and Blei, 2007; Lacoste-Julien et al., 2008). E.g., Labeled LDA (Ramage et al., 2009) constrains Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model through one-to-one correspondence between LDA's latent topics and document labels. While they improve the predictive ability of unsupervised topic models, they require a massive set of annotated documents to be effective. The second type of supervision comes in the form of seed words where for each user-interested topic, a set of represented keywords are provided to guide the topic generation process (Jagarlamudi et al., 2012; Eshima et al., 2020; Harandizadeh et al., 2022; Gallagher et al., 2017). More specifically, the seed guided topic models make the topics converge in the direction of given seed words. Although they make the topics converge in the direction of user-desired seed words, they require the seed words to appear in the corpus vocabulary, which may be impractical in many cases. Finally, there is an approach called category-guided topic mining (Meng et al., 2020a) (CatE), which considers the topics' surface names as the only supervision for mining user-interested discriminative topics. However, it also assumes that the surface name of each topic appears in the corpus. Moreover, none of the above work does impose requirements of making topics comparable over multiple corpora. Moreover, because of strict discrimination assumptions in CatE, they often suffer from mining too specific words to represent topics that are very hard to interpret.

B Implementation Details

There are some parameters our model we have to set. We use the 300-dimensional pre-trained word embedding from Spacy (Honnibal and Montani, 2017). The dimension of hidden layers in MLP is set to 300. The learning rate is set to 0.005. We use different epochs based on different datasets. We mainly use 150 for news domain datasets and 100 for other cases. The regularization parameters λ_β and λ_θ are set to 20 and 35, respectively, for all datasets. We follow the standard procedure for preprocessing text before applying all the models. For example, we remove stopwords; the words appear more than 0.70 of the whole corpus, and infrequent words appear less than ten times. For baselines, we used their default settings.

C Additional Text Classification Baselines

- SupBERTCLs: It trains a BERT-based supervised text classifier [6] on labeled reference corpus and uses that to predict labels for documents in target corpus.
- Zero-ShotCLs: It predicts the topic of each document out of given surface names $\mathcal{C}$ using a Zero-Shot classification (Yin et al., 2019) method based on a pre-trained textual entailment model[7]
- WeSTCLs: This is a weakly-supervised neural text classification method (Meng et al., 2018) that

[6] https://huggingface.co/docs/transformers/tasks/sequence_classification

[7] https://huggingface.co/joeddav/bart-large-mnli-yahoo-answers

Dec. 2000	court, election, vote, bush, president, florida, law, judge, justice, government, supreme, party, clinton, official, case, presidential, decision, ballot, republican, democratic
Sep. 2001	united, attack, war, president, government, official, american, bush, terrorist, military, country, terrorism, national, afghanistan, vote, election, administration, leader, political, muslim
USA	republican, vote, campaign, election, president, party, governor, senator, candidate, political, mayor, democratic, clinton, bush, voter, democrat, democrats, giuliani, senate, race
Britain	minister, president, ireland, party, government, political, official, prime, irish, northern, clinton, leader, peace, blair, union, thatcher, republican, war, election, sinn

Table 6: Top 20 words in politics topic on different contexts

can classify a text document based on given class names or seed words. We use the two variants: (1) WeSTCLs (Names) uses the topic names $\mathcal{C}$; (2) WeSTCLs (Seeds) uses seed words.

- LOTCls: This is a language model-based text classification method (Meng et al., 2020b) that uses label names as supervision. As supervision, we provide the given topic surface names $\mathcal{C}$.

D Additional Result on Corpus Comparison

Table 6 shows qualitative results of corpus comparison. It shows the top 20 words for each corpus generated by our ECTM model, each representing a context for the topic of politics.