Continual Neural Topic Model

Charu Karakkaparambil James, Waleed Mustafa,
Marcio Monteiro , Marius Kloft, Sophie Fellenz
RPTU University Kaiserslautern-Landau
Kaiserslautern, Germany
charu@cs.uni-kl.de, mustafa@cs.uni-kl.de,
marcio.monteiro@cs.rptu.de, kloft@cs.uni-kl.de, fellenz@cs.uni-kl.de

Abstract

In continual learning, our aim is to learn a new task without forgetting what was learned previously. In topic models, this translates to learning new topic models without forgetting previously learned topics. Previous work either considered Dynamic Topic Models (DTMs), which learn the evolution of topics based on the entire training corpus at once, or Online Topic Models, which are updated continuously based on new data but do not have long-term memory. To fill this gap, we propose the Continual Neural Topic Model (CoNTM), which continuously learns topic models at subsequent time steps without forgetting what was previously learned. This is achieved using a global prior distribution that is continuously updated. In our experiments, CoNTM consistently outperformed the dynamic topic model in terms of topic quality and predictive perplexity while being able to capture topic changes online. The analysis reveals that CoNTM can learn more diverse topics and better capture temporal changes than existing methods.

1 Introduction

Topic models are used to discover the hidden thematic structure in a collection of documents. These models are particularly useful in Natural Language Processing (NLP), supporting a wide range of applications, including information extraction, text clustering, summarization, sentiment analysis, content recommendation, opinion/event mining, and trend analysis (Tuan et al., 2020; Subramani et al., 2018; Nguyen et al., 2021; Wang and Mengoni, 2020; Molenaar et al., 2024; Wang and Blei, 2011; Avasthi et al., 2022; Churchill and Singh, 2022b). A popular topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which represents each document as a collection of topics, with each topic being a distribution over words.

However, document collections are usually not recorded at a singular instance in time. Document collections may span many months or years, during which the topics (and their word distributions) may change. Standard LDA is not designed to handle such dynamic changes.

Dynamic topic models, such as Dynamic LDA (Blei and Lafferty, 2006), Dynamic Embedded Topic Model (Dieng et al., 2019), and Dynamic BERtopic (Grootendorst, 2022), address this issue by capturing the evolution of topics over time.

Despite their advancements, these models share a limitation: *they require the entire corpus to be available from the start.* In real-world applications, new data is generated every day, and it should ideally be processed online in real-time. For example, as data are continuously streamed, there is significant potential in monitoring current topics of interest, detecting emerging trends in social media (Sasaki et al., 2014), analyzing consumer purchase behavior (Iwata et al., 2009), and tracking urban geo-topics (Yao and Wang, 2020).

To address the need of real-time processing, online topic models have been introduced (AlSumait et al., 2008; Iwata et al., 2010; Zhang et al., 2013), which adapt to data arriving sequentially.

However, online topic models lack long-term memory and tend to forget previously acquired topic knowledge over time. This is where continual learning becomes relevant. In continual learning, new sub-problems are learned over time without forgetting what was previously learned.

We propose the *Continual Neural Topic Model (CoNTM)*, which uses a global prior distribution to store information over time, while local models capture patterns inherent in the current time step. The proposed model effectively captures global thematic patterns and tracks their temporal evolution over locally defined subsets of the data. The CoNTM maintains high topic quality and low predictive perplexity without losing previously learned information. This is achieved as each time step depends on the global prior, ensuring consistency and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6636–6658
March 24-29, 2026 ©2026 Association for Computational Linguistics

coherence over time.

The contributions of the paper are as follows:

- We introduce CoNTM, a continual neural topic model to train a sequence of topic models without forgetting what was previously learned.

- We introduce the CoNTM algorithm, which incrementally updates global topics at each time step t, thereby capturing topic dependencies from the previous time step.

- Through experiments on six diverse datasets, we observe that CoNTM outperforms state-of-the-art DTMs regarding topic quality and predictive perplexity.

- Unlike traditional models that require access to all data in advance, we show CoNTM maintains good qualitative performance even when training in a data stream.

We discuss related work in Section 2. The proposed continual modeling methodology is described in more detail in Section 3. Section 4 describes the evaluation measures, datasets, and model settings and presents qualitative and quantitative results. Section 5 provides the conclusion.

2 Related Work

Tracking the evolution of topics over time has so far been addressed in the research areas of dynamic topic models and online topic models, both of which we review below. We further discuss relevant work in the area of continual learning.

Dynamic Topic Models Dynamic topic models assume that the complete corpus is available for training. Rahimi et al. (2023) differentiates between probabilistic dynamic topic models (PDTMs) and algorithmic dynamic topic models (ADTMs).

PDTMs are based on generative assumptions. Previous work on PDTMs includes Dynamic LDA (Blei and Lafferty, 2006), the Dynamic Embedded Topic model (DETM) (Dieng et al., 2019), Dynamic Structured Neural Topic Model with Self-Attention Mechanism (Miyamoto et al., 2023), Dynamic Noiseless LDA (DNLDA) (Churchill and Singh, 2022a), modeling discrete dynamic topics (Bahrainian et al., 2017) and Continuous Time Dynamic Topic Models (Wang et al., 2012). DLDA, a probabilistic model, is not based on neural networks and is not scalable to large datasets. DETM combines latent Dirichlet allocation (DLDA) and word embeddings. In DETM, each word is modeled with a categorical distribution parameterized by the inner product of the word embedding and the embedded representation of the topic at each step in time. We compare our model with these baselines in the experiments section. However, unlike our online model, they are batch models and cannot be updated in a continuous data stream. Another PDTM by Tomasi et al. (2022) improves rare word inclusion using the correlation-based method and amortized variational inference, making it more efficient for large vocabularies.

ADTMs do not assume a document generation process, but cluster document embeddings and extract topic words using heuristic methods. Unlike PDTMs, which jointly learn topic clusters and embeddings, ADTMs separate these steps. They excel at short texts like Tweets but face challenges with domain-specific corpora due to reliance on pretrained embeddings. Examples include the BERTopic algorithm (Grootendorst, 2022) using the BERT language model (Devlin et al., 2018), ANTM (Rahimi et al., 2023), Dynamite (Balepur et al., 2023), CFDTM (Wu et al., 2024a) uses contrastive learning techniques, and Dynamic BERTopic, which extends BERTopic using c-TF-IDF. Other models by Eklund et al. (2022); Gao et al. (2022); Boutaleb et al. (2024) also cluster embeddings from pretrained language models (e.g., BERT, GPT) to track topics.

All these models rely on pretrained embeddings. In contrast, our model learns domain-specific topics even if the above-mentioned pretrained model does not include the target domain.

Online Topic Models Online topic models are updated as new data arrives. Previous work includes Online LDA (OLDA) (AlSumait et al., 2008), which incrementally adds new data to the current model rather than requiring access to previous data. The sparse online topic model (Zhang et al., 2013) uses sparsity-inducing regularization to control the sparsity of latent semantic patterns and employs online algorithms to learn the topical dictionary. The multiscale dynamic topic model (Iwata et al., 2010) incrementally updates the model at each epoch using the newly obtained set of documents and the multiscale model from the previous epoch. Banerjee and Basu (2007) provide a study on batch and online unsupervised learning. This work is based on statistical topic models and has not yet been adapted to neural network-based topic models with long-term memory capability.

Neural network-based topic models have been studied by Srivastava and Sutton (2017); Miao et al. (2016); Burkhardt and Kramer (2019); Dieng et al. (2020b); Bianchi et al. (2020a); Srivastava and Sutton (2017); Reimers and Gurevych (2019); Grootendorst (2022); Panwar et al. (2021); Wu et al. (2024b). In contrast to our model, these neural topic models do not capture the evolution of topics over time.

Continual Learning Continual learning (Hadsell et al., 2020; Zenke et al., 2017; Xu and Zhu, 2018) aims at developing systems that can continuously learn and adapt to new data or tasks over time without forgetting previously learned information. Gupta et al. (2020) use continual learning for topic models on sparse data, where small collections of documents often lead to incoherent topics. In their work, the authors use multiple-shot task learning with multiple datasets from different domains.

In contrast, our model applies continual learning to learn new topics within a domain-specific dataset without forgetting previously learned topics.

Importantly, task-based continual learning methods assume a sequence of discrete, well-separated tasks, often with explicit task boundaries or domain shifts. In contrast, topic evolution in temporal corpora exhibits gradual, within-domain distributional drift, where topics persist, transform, and overlap across time without clear task demarcations. As a result, task-conditioned CL strategies such as task-specific regularization or replay are not directly applicable to our setting.

3 Methodology

This section presents the proposed Continual Neural Topic Model (CoNTM).

3.1 Preliminaries

The Dirichlet Variational AutoEncoder (DVAE) forms the foundation of the proposed method. In DVAE, a variational autoencoder (VAE) generates the document-topic distribution using an encoder network parameterized by θ. This encoder network effectively captures the thematic structure of the documents by mapping them onto a latent topic space. This means that each document has a document-topic distribution $z \sim \text{Dirichlet}(\alpha)$ with a Dirichlet prior. Using a Dirichlet prior is essential as it encourages sparsity in the topic distributions, thereby enhancing the interpretability of the topics. On the other hand, the topic-word distribution, which represents the probability of words given a particular topic, is represented by a decoder network that reconstructs the input documents.

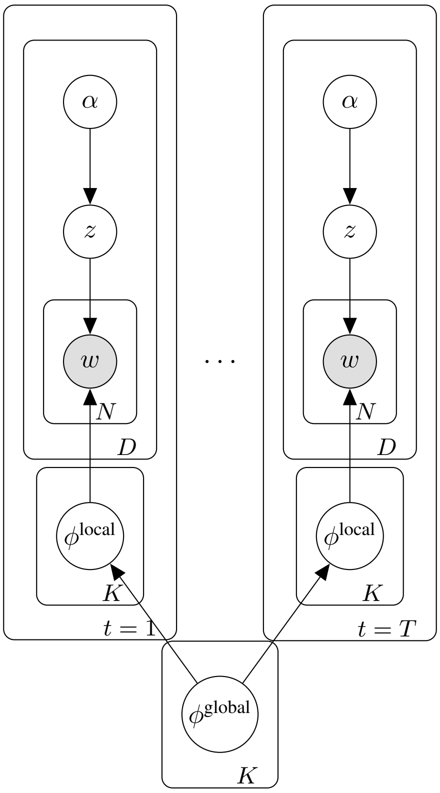


Figure 1: The Graphical Model of our approach. The local models of each time slice t are connected by the global parameters ϕ^{global}, which captures topic dependencies from the previous time step.

3.2 Continual Neural Topic Model (CoNTM)

In CoNTM, we model documents as arriving in continuous time slices, with each slice characterized by slightly varying topics. These topics are interconnected through a global topic set, allowing for minor temporal adjustments to the global topics at each time step t. This approach implicitly captures topic dependencies from the previous time step via the global distribution while offering the advantages of a reduced number of parameters and increased model flexibility as compared to approaches that explicitly model each transition.

Formally, a document at time t is modeled as a mixture of local topics $\phi_t^{\text{local}} = (\phi_{t,1}^{\text{local}}, \dots, \phi_{t,K}^{\text{local}})$, where each topic $\phi_{t,k}^{\text{local}}$ is a probability distribution over the vocabulary. We further assume that the local topics are derived from a set of global topics ϕ^{global} such that $\phi_t^{\text{local}} = g(\phi^{\text{global}}, \Delta\phi_t^{\text{local}})$,

Algorithm 1 data generative algorithm

```
Input: Number of time slices $T$, documents $\mathbf{w}$
Output: Local Topics $\{\phi_t^{\text{local}}\}_{t=1}^{T}$
for $t = 1 \ldots T$ do
    $\phi_t^{\text{local}} \leftarrow g(\phi^{\text{global}}, \Delta\phi_t^{\text{local}})$
    for each document $\mathbf{w}$ do
        Draw a document topic distribution:
            $z \sim \text{Dirichlet}(\alpha)$
        for each word index $n$ do
            Draw a word:
                $\mathbf{w}_n \sim \text{Multinomial}(1, \phi_t^{\text{local}} z)$
        end for
    end for
end for
```

where g is a transformation function that applies a perturbation $\Delta\phi_t^{\text{local}}$ to the global topics to obtain the local topics. This assumption ensures that local topics remain consistent over time. The graphical model is shown in Figure 1.

3.2.1 Generative Process

The algorithm 1 outlines the assumed generative process of our data. According to this generative process, the marginal distribution of a document $\mathbf{w}$ at time t is given by

$$p(\,\mathbf{w}\,|\alpha, \phi^{\text{global}}, \Delta\phi_t^{\text{local}}) = \int_z p(z \mid \alpha) \prod_{n=1}^{N} p(\mathbf{w}_n \mid z, g(\phi^{\text{global}}, \Delta\phi_t^{\text{local}}))\, dz$$

where $p(z \mid \alpha)$ is the Dirichlet distribution over topic proportions, and $p\left(\mathbf{w}_n \mid z, g\left(\phi^{\text{global}}, \Delta\phi_t^{\text{local}}\right)\right)$ is the multinomial distribution over words given the local topics.

3.2.2 Variational Inference

The posterior inference over the parameters z is intractable (Srivastava and Sutton, 2017). We thus resort to the DVAE framework (Burkhardt and Kramer, 2019). That is, we assume a variational distribution $q_\theta(z)$ on the random variable z parameterized by a free parameter θ, which is learned by maximizing the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta, \mathbf{w}\,|\phi^{\text{global}}, \Delta\phi_t^{\text{local}}) = -\text{KL}(q_\theta(z) \,\|\, p(z \mid \alpha)) + \mathbb{E}_{q_\theta(z)}[\log p(\mathbf{w} \mid z, \phi^{\text{global}}, \Delta\phi_t^{\text{local}})].$$

Algorithm 2 CoNTM algorithm

```
Input: Number of time slices $T$, stream of documents $\{\mathbf{w}^{i,1}\}_{i=1}^{n_1}, \cdots, \{\mathbf{w}^{i,T}\}_{i=1}^{n_T}$, number of Topics $K$, and number of training steps $J$
Output: Topics: $\{\hat{\phi}_t^{\text{local}}\}_{t=1}^{T}$, $\hat{\phi}^{\text{global}}$
for $t = 1 \ldots T$ do
    Initialize $\theta$
    Initialize $\Delta\hat{\phi}_t^{\text{local}}$
    $\{\mathbf{w}^{i,t}\}_{i=1}^{n_t} \leftarrow$ documents arrived at time $t$
    for $j = 1 \ldots J$ do
        Update $\theta$ by the gradient
            $\nabla_\theta \sum_{i=1}^{n_t} \mathcal{L}(\theta, \mathbf{w}^{i,t} \mid \hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}})$
        Update $\Delta\hat{\phi}_t^{\text{local}}$ by the gradient
            $\nabla_{\Delta\hat{\phi}_t^{\text{local}}} \sum_{i=1}^{n_t} \mathcal{L}(\theta, \mathbf{w}^{i,t} \mid \hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}})$
    end for
    Set $\hat{\phi}_t^{\text{local}} \leftarrow \hat{\phi}^{\text{global}} + \Delta\hat{\phi}_t^{\text{local}}$
    Set $\hat{\phi}^{\text{global}} \leftarrow (1 - \rho_t)\hat{\phi}^{\text{global}} + \rho_t\hat{\phi}_t^{\text{local}}$
end for
```

The form of $q_\theta(z)$ is characterized by an encoder network $\alpha_\theta(\mathbf{w})$ parameterized by θ. Specifically, the variational distribution is defined as $\text{Dirichlet}(\alpha_\theta(\mathbf{w}))$.

Optimizing the ELBO with respect to ϕ_t^{local} and ϕ^{global} can be challenging due to the simplex constraint ($\sum_w \phi_{k,w} = 1$). Therefore, we follow the DVAE topic model approach by introducing unconstrained variables $\Delta\hat{\phi}_t^{\text{local}}$ and $\hat{\phi}^{\text{global}}$. Using these variables, we define the probability of the n-th word $\mathbf{w}_n$ in a document as follows:

$$p(\mathbf{w}_n = w \mid z, \hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}}) = [\sigma(g(\hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}}) \cdot z)]_w,$$

where σ denotes the softmax function. Notably, the softmax normalization is performed after mixing the resulting local topics with the topic weights z. This approach is shown to enhance the model's expressive power by treating the word distribution as a product of experts (Srivastava and Sutton, 2017).

3.2.3 Global Topic Parameters

We now turn our attention to the function g, which links the local to the global topic parameters. While our approach accommodates a general form for g, we assume a simple form:

$$\hat{\phi}_t^{\text{local}} = g(\hat{\phi}^{\text{global}}, \Delta\hat{\phi}_t^{\text{local}}) := \hat{\phi}^{\text{global}} + \Delta\hat{\phi}_t^{\text{local}}.$$

We further introduce the constraint $\sum_{t=1}^{T} \Delta\hat{\phi}_t^{\text{local}} = 0$. This can be interpreted as $\Delta\hat{\phi}_t^{\text{local}}$ being realized from a centered probability distribution. The choice of g and the constraint on $\Delta\hat{\phi}_t^{\text{local}}$ has the advantage that

$$\hat{\phi}^{\text{global}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\phi}_t^{\text{local}}.$$

Thus, inference is simplified by only computing the local topics $\hat{\phi}_t^{\text{local}}$ and obtaining the global topics by taking their average. Since we are interested in continual learning, we replace taking the average with a running average of the local topics, that is

$$\hat{\phi}^{\text{global}} = \frac{t-1}{t}\hat{\phi}^{\text{global}} + \frac{1}{t}\hat{\phi}_t^{\text{local}}.$$

To control the amount of updates to $\hat{\phi}^{\text{global}}$ over time, we further replace $\frac{t-1}{t}$ (resp. $\frac{1}{t}$) by $1-\rho_t$ (resp. ρ_t) for $\rho_t \in (0,1)$. For instance, we can set $\rho_t = \frac{1}{(\tau_0+t)^\kappa}$, where $\kappa \in (0.5, 1]$ and $\tau_0 \geq 0$ as in Hoffman et al. (2010). Here, κ controls the rate of forgetting the old estimate of the global topics, while τ_0 slows down the updates in the early steps. Selecting proper values for τ_0 and κ ensures that the influence of local updates decreases over time, allowing the model to converge to a global topic-word distribution. Algorithm 2 summarizes the learning procedure.

4 Experiments

This section compares the proposed model with five baselines, presenting quantitative results and a qualitative analysis of topic evolution.

4.1 Evaluation of Dynamic Topic Models

We use topic coherence (TC) and topic diversity (TD) to evaluate the dynamic topic model. A commonly used TC metric is the Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009; Lau et al., 2014; Dieng et al., 2019, 2020a; Bianchi et al., 2020b; Nagda et al., 2021). A temporal reference corpus was utilized to determine the NPMI score for the topics. This means that the NPMI score for a topic at a specific timestamp is calculated using the reference corpus available up to that point in time. Burkhardt and Kramer (2019) proposed Topic Redundancy (TR), a measure that calculates the average occurrences of a top word in other topics. The topic diversity is calculated as $TD = 1 - TR$. The redundancy for topic k is given below:

$$TR(k) = \frac{1}{K-1}\sum_{i=1}^{N}\sum_{j\neq k} P\left(w_{ik}, j\right).$$

Here, $P\left(w_{ik}, j\right)$ equals one if the ith word of topic k, w_{ik}, occurs in topic j and otherwise zero. $K-1$ is the number of topics excluding the current topic.

To ensure that the topic quality is not affected by the occurrence of too few or too many topics, it is normalized based on the total number of topics in each timestamp (Rahimi et al., 2023):

$$\text{TQ} = \frac{1}{k}\sum_{i=0}^{k-1} \text{TC}_i \times \text{TD}_i \times \frac{T_i}{\text{T}_i^{\max}}.$$

Here TC_i and TD_i represent topic coherence and diversity in timestamp i. Additionally, T_i represents the number of topics within timestamp i, and $\text{T}_i^{\max}$ indicates the highest number of topics observed across all years. To track topic changes over time, we use the recently proposed Temporal Topic Smoothness (TTS) measure (Karakkaparambil James et al., 2024). The TTS indicates whether the topic transition is abrupt or gradual.

Finally, we also use predictive perplexity (PPL) on unseen future timestamps (Wang et al., 2012), a standard evaluation measure for assessing the performance of probabilistic language models. Lower perplexity values indicate better predictive model performance.

4.2 Datasets

Our study is conducted on six widely recognized datasets within the field (see Table 2 in Appendix B). The first is a collection of articles from the New York Times (Sandhaus, 2008), covering a span of two decades, specifically from 1987 to 2007. The second, the UN corpus (Jankin Mikhaylov et al., 2017a), encompasses a temporal range of five decades, extending from 1970 to 2020, and comprises statements from the general debates over this period. The third dataset under study is the NIPS corpus (Swami, 2020), which includes the entirety of the NIPS conference published between 1987 and 2019. The fourth dataset contains around 14,000 tweets from @NASA's Twitter account, spanning a period of over four years, from 2018 to 2022. The fifth dataset includes a collection of

	CoNTM (ours)			**DETM**			**DLDA**			**DBERTopic**			**DNLDA**			**CFDTM**		
Dataset	TC ↑	TD ↑	TQ ↑	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ
NIPS	.087	.969	.084	-.009	.970	-.009	.097	.980	.095	.058	.343	.032	.023	.997	.022	-.112	.401	-.045
NYT	.174	.991	.173	.137	.987	.135	.122	.939	.116	.117	.797	.082	.069	.997	.069	.155	.756	.117
UN	.085	.863	.074	-.045	.958	-.043	.096	.942	.091	.057	.515	.024	.034	.994	.034	.093	.768	.073
Tweets	-.003	.907	-.006	-.008	.982	-.008	.036	.965	.019	.114	.912	.096	-.135	.994	-.135	-.287	.487	-.139
Arxiv	.102	.974	.099	.069	.963	.067	.084	.966	.082	.064	.962	.059	-.004	.996	-.004	.096	.830	.080
DBLP	.118	.969	.115	.073	.966	.071	.101	.958	.097	.093	.951	.086	.041	.998	.040	.124	.874	.108
Av. Rank	**2.0**	3.1	**1.6**	4.5	2.8	4.3	2.3	3.1	2.1	3.6	5.1	3.8	5.3	**1.0**	5.1	3.1	5.6	3.8

Table 1: Comparison of six topic models (CoNTM, DETM, DLDA, DBERTopic, DNLDA, and CFDTM) on six datasets with 50 topics. Metrics: topic coherence (TC), topic diversity (TD), and topic quality (TQ), averaged over three runs. Best values per dataset in **bold**.

16,000 documents, consisting of arXiv titles and abstracts (arXiv.org submitters, 2023), covering the years 2012 to 2024. The final dataset consists of the DBLP archive (Ley, 2002), which includes 168,000 scientific articles (titles and abstracts) published between 2000 and 2020. Additionally, the details of the preprocessing steps can be found in Appendix B.

4.3 Models

The CoNTM model is evaluated alongside five distinct baseline models: DETM (Dieng et al., 2019), DLDA (Blei and Lafferty, 2006), Dynamic BERTopic (Grootendorst, 2022), DNLDA (Churchill and Singh, 2022a), and CFDTM (Wu et al., 2024a). DLDA and DNLDA offer more traditional probabilistic approaches with a focus on topic evolution, while DETM and Dynamic BERTopic leverage embeddings to capture semantic changes. For the CoNTM, a learning rate of 0.01 is adopted. The model employs the Adam optimizer and divides the dataset into an 80% training set, a 10% validation set, and a 10% test set. Refer to Appendix A for hyperparameter settings of other models and Appendix N for CoNTM experiments with varying learning rates and optimizers. For the experiment, a κ value of 0.7 and a τ value of 1 were used, and the sensitivity analysis is presented in Appendix F. The number of topics used across all models in this study is uniformly set to 50. Experimental results for the models configured with 20 topics are shown in Appendix E, demonstrating stability across varying topic sizes. Additional experiments on generative predictive perplexity are presented in Appendix L.

4.4 Quantitative Results

This section compares the CoNTM model to four different baselines, focusing on coherence, diversity, smoothness, and temporal aspects while highlighting model strengths on domain-specific datasets.

Topic Coherence vs Diversity: After analyzing topic coherence, diversity, and quality, we found that our CoNTM model exhibits good topic quality with an average rank of 1.6 among the five models (see Table 1) and maintains good topic coherence score (see Figure 2) on large datasets. The value presented in the table is the average of three random seed values. The average rank for each model is computed by averaging its ranks across all datasets. CoNTM, DLDA, and DBERTopic showed moderate topic coherence, whereas DETM and DNLDA presented varied results (see Table 1). Additionally, as shown in Figure 2, the CoNTM model achieved good topic diversity and coherence on the NYT, DBLP, and Arxiv datasets, which contained sufficiently large documents. The alternate datasets, NIPS, UN, and Tweets, are described in Appendix I. Models in the top-right corner exhibit good topic coherence and good topic diversity, indicating strong performance.

For smaller datasets, such as Tweets, which consist of approximately 9,703 documents, the Dynamic BERTopic model exhibits good performance based on topic quality. This model uses pre-trained word embeddings, significantly improving the topic quality compared to other models in small datasets. However, it is observed that as the size of the dataset increases (see Table 2), the performance of the Dynamic BERTopic model tends to decline, ex-

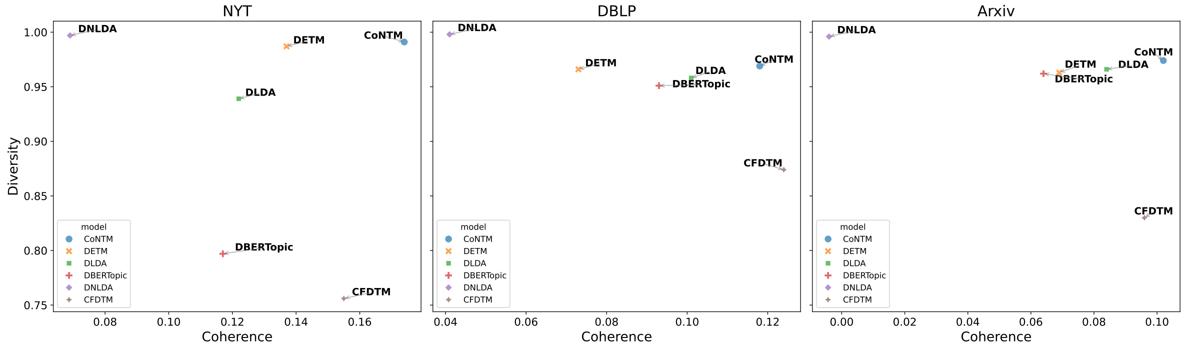


Figure 2: The figure shows the model performance quantitatively for the NYT, BDLP, and Arxiv datasets. The top-right corner indicates that the model achieves high topic quality and low predictive perplexity. Our model (CoNTM) outperformed the other models for these datasets in terms of both coherence and diversity.

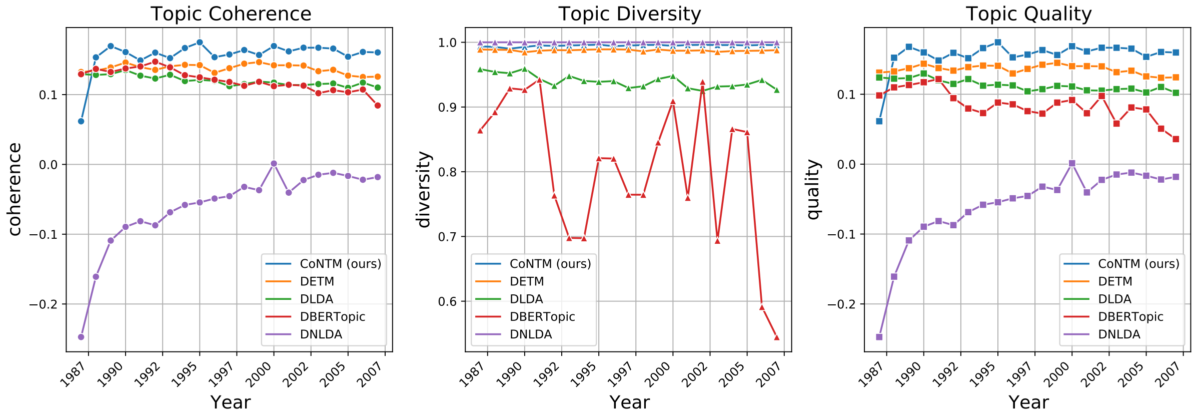


Figure 3: The figure illustrates how coherence, diversity, and topic quality change over the years for the NYT Dataset. Our model shows increasing topic coherence over the year and outperforms other models. Both the DETM and DLDA show stable coherence, while the Dynamic BERT model shows a decrease in coherence over the year.

cept for Tweets. Furthermore, for the NIPS dataset, the CoNTM model exhibits moderate topic quality compared to other models. This is due to the insufficient documents available at the early timestamp for the NIPS dataset. This affects topic quality in subsequent timestamps, as our model is continual.

We can also observe that for Dynamic BERTopic, the topic quality is lower on domain-specific datasets like the UN compared to general datasets. This may be because large language models are not trained on domain-specific datasets, leading to a decline in topic quality when using pre-trained word embeddings.

In summary, the CoNTM model consistently demonstrates strong topic quality, highlighting its effectiveness in extracting coherent and meaningful topics from the data. Notably, this performance is achieved without the assumption that all data are available from the start. Instead, CoNTM is capable of handling scenarios where data arrive incrementally.

4.5 Evaluation on Downstream Tasks

Figure 5 compares six topic models: DETM, DLDA, DBERTopic, DNLDA, CFDTM, and the proposed CoNTM on the NYT dataset (five labels) across two tasks, *text classification* and *document clustering*, using Accuracy and F1 Score.

For both downstream tasks, we evaluated the learned topic representations (θ) using classification and clustering. In the classification setup, a Logistic Regression model with the `lbfgs` solver (max 5000 iterations) was trained on document topic distributions from the training set and tested on the held-out data, measuring performance via Accuracy and macro-averaged F1 Score. For clustering, the K-Means algorithm was applied in an unsupervised manner with the number of clusters matching the ground-truth classes. Predicted clusters were aligned with true labels using the Hungarian algorithm on the contingency matrix, and final Accuracy and macro-F1 scores were computed to evaluate how well the topic space preserves the

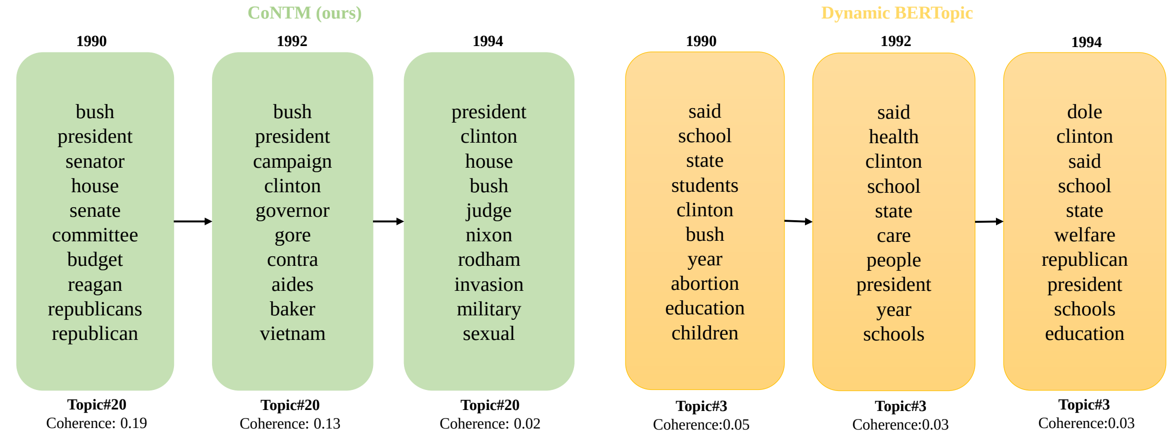

CoNTM (ours)
Dynamic BERTopic
1990
1992
1994
bush president senator house senate committee budget reagan republicans republican
bush president campaign clinton governor gore contra aides baker vietnam
president clinton house bush judge nixon rodham invasion military sexual
Topic#20
Coherence: 0.19
Topic#20
Coherence: 0.13
Topic#20
Coherence: 0.02
1990
1992
1994
said school state students clinton bush year abortion education children
said health clinton school state care people president year schools
dole clinton said school state welfare republican president schools education
Topic#3
Coherence:0.05
Topic#3
Coherence:0.03
Topic#3
Coherence:0.03

Figure 4: The figure shows the progression of two topics from CoNTM and Dynamic BERTopic in the NYT dataset from 1990 to 1994. Our focus is on the US election in 1992, where *Clinton* was elected as president. CoNTM demonstrates better topic coherence, capturing *Clinton* as a top word in 1992, while Dynamic BERTopic shows little change over time.

underlying semantic structure.

Overall, CoNTM achieves the best performance across both tasks, reaching an Accuracy and F1 Score of 0.894 in classification, surpassing CFDTM (0.880, 0.881) and DBERTopic (0.808, 0.805), and leading in clustering with 0.705 Accuracy and 0.707 F1, notably outperforming CFDTM (0.313, 0.240) and DETM (0.435, 0.425).

These results demonstrate that CoNTM effectively captures semantic and structural information within topics, enabling both accurate classification and coherent cluster formation. The consistent performance across tasks highlights the model's ability to generate high-quality, discriminative topic representations in the NYT corpus.

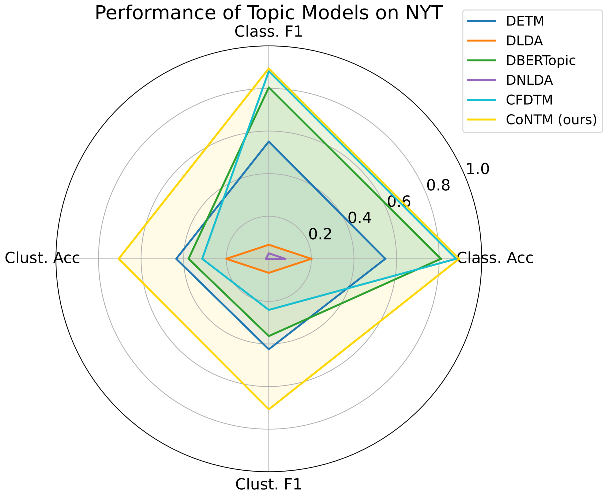


Figure 5: Radar plot showing topic model performance on the NYT dataset across four metrics: Classification Accuracy, Classification F1, Clustering Accuracy, and Clustering F1. Models nearer the outer boundary perform better, with CoNTM (ours) consistently outperforming all baselines.

4.6 Qualitative Results

This section qualitatively analyzes the "politics" topic as it evolves over time. Additional analysis is shown in Appendices C and D, indicating that CoNTM and DETM produce more coherent topics for the NYT dataset while tracking topic evolution.

Evolving Topics: CoNTM effectively captures evolving topics, as shown in Figure 4, highlighting the 1992 election of *Clinton*. Key topic words during this timestamp include "bush", "president", "campaign", and "clinton". In contrast, Dynamic BERTopics shows little change over time, not adequately representing the events per year. Figure 19 in Appendix K additionally illustrates the evolving topic "politics" with the word probability on the y-axis and timestamp on the x-axis. We observe an increase in the word probability of *Clinton* in 1997, which corresponds to his inauguration for a second term as the 42nd President of the United States. His final years in office were from 1999 to 2000. Notably, while CoNTM captures major political events such as the 1992 U.S. election, the transition remains gradual across adjacent timestamps, reflecting the model's bias toward smooth evolution rather than instantaneous topic birth. See Appendix K for an additional example from the UN dataset.

Topic Coherence: The topics generated by CoNTM exhibit high coherence (Figure 4), with clear connections among the words. For instance, the transition to *Clinton's* 1992 election highlights political themes. This suggests that CoNTM effectively captures topic changes over time, maintaining a strong thematic connection. While Dynamic

BERTopics can capture a broad range of topics, the coherence within each topic is lower than CoNTM. In conclusion, CoNTM excels in detecting emerging topics with good topic coherence.

5 Conclusion

We have presented a novel Continual Neural Topic Model (CoNTM), a DVAE-based method for topic modeling that continuously learns evolving topics without forgetting previously learned information. We evaluate CoNTM using datasets from various domains, including news, politics, science, and Tweets, through both quantitative and qualitative analysis. Furthermore, CoNTM demonstrates the ability to track temporal evolution in real-time sequential data. Notably, CoNTM outperforms the DTM models when dealing with large datasets, even though the CoNTM model does not assume all data to be available from the start.

In the future, we want to extend our model to handle datasets of varying sizes by incorporating word embeddings. Also an interesting direction for future work would be to integrate nonparametric mechanisms into the global topic update, enabling adaptive topic allocation while preserving continual learning behavior.

Limitations

The Continual Neural Topic Model, while offering significant advances in topic modeling, especially in its ability to capture the evolution of a topic over time, comes with some limitations. First, these models often require a large amount of data to train effectively. In scenarios where data is sparse or where topics evolve rapidly, the model might struggle to learn meaningful patterns. Second, while these models are designed to capture topic evolution over time, they may not always accurately reflect rapid shifts in topics, particularly in fast-changing domains like social media or news.

Third, evaluating the quality of topics and their temporal evolution remains a challenge, as traditional topic coherence metrics may not fully capture the semantic shifts over time, making reliable assessment and comparison with static models difficult. The fourth limitation is that it assumes a fixed number of topics over time and does not explicitly model topic birth or death. The CoNTM instead focuses on stable topic identity and smooth semantic evolution, trading flexibility for robustness and interpretability in streaming settings.

Acknowledgements

The authors acknowledge support by the BMFTR award 01IS24071A, DFG through FOR 5359 (ID 459419731), TRR 375 (ID 511263698), SPP 2298 (Id 441826958), and SPP 2331 (441958259, 553345933, 466468799), and the Carl-Zeiss Foundation through the initiatives AI-Care and Process Engineering 4.0.

References

Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. 2008. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 eighth IEEE international conference on data mining*, pages 3–12. IEEE.

arXiv.org submitters. 2023. arxiv dataset.

Sandhya Avasthi, Ritu Chauhan, and Debi Prasanna Acharjya. 2022. Topic modeling techniques for text mining over a large-scale scientific and biomedical text corpus. *International Journal of Ambient Computing and Intelligence (IJACI)*, 13(1):1–18.

Seyed Ali Bahrainian, Ida Mele, and Fabio Crestani. 2017. Modeling discrete dynamic topics. In *Proceedings of the symposium on applied computing*, pages 858–865.

Nishant Balepur, Shivam Agarwal, Karthik Venkat Ramanan, Susik Yoon, Diyi Yang, and Jiawei Han. 2023. Dynamite: Discovering explosive topic evolutions with user guidance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 194–217.

Arindam Banerjee and Sugato Basu. 2007. Topic models over text streams: A study of batch and online unsupervised learning. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 431–436. SIAM.

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, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

D. M. Blei and J. D. Lafferty. 2006. Dynamic topic models. *In International Conference on Machine Learning*.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Allaa Boutaleb, Jerome Picault, and Guillaume Grosjean. 2024. Bertrend: Neural topic modeling for emerging trends detection. *arXiv preprint arXiv:2411.05930*.

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

Rob Churchill and Lisa Singh. 2022a. Dynamic topic-noise models for social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 429–443. Springer.

Rob Churchill and Lisa Singh. 2022b. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

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 J. R. Ruiz, and David M. Blei. 2020a. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

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

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The dynamic embedded topic model. *ArXiv*, abs/1907.05545.

Anton Eklund, Mona Forsman, and Frank Drewes. 2022. Dynamic topic modeling by clustering embeddings from pretrained language models: A research proposal. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 84–91, Online. Association for Computational Linguistics.

Q. Gao, X. Huang, K. Dong, Z. Liang, and J. Wu. 2022. Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. *Scientometrics*, 127(3):1543–1563.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

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.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. 2020. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23.

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

Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. 2009. Topic tracking model for analyzing consumer purchase behavior. In *Twenty-First international joint conference on artificial intelligence*.

Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. 2010. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 663–672.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. 2017a. United Nations General Debate Corpus.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. 2017b. United Nations General Debate Corpus.

Charu Karakkaparambil James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft, and Sophie Fellenz. 2024. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–176, Bangkok, Thailand. Association for Computational Linguistics.

Oliver Kramer and Oliver Kramer. 2016. Scikit-learn. *Machine learning for evolution strategies*, pages 45–53.

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.

Michael Ley. 2002. The dblp computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*, pages 1–10. Springer.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, and Ichiro Sakata. 2023. Dynamic structured neural topic model with self-attention mechanism. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5916–5930, Toronto, Canada. Association for Computational Linguistics.

Annika Molenaar, Dickson Lukose, Linda Brennan, Eva L Jenkins, and Tracy A McCaffrey. 2024. Using natural language processing to explore social media opinions on food security: Sentiment analysis and topic modeling study. *Journal of Medical Internet Research*, 26:e47826.

Marcio Monteiro, Charu Karakkaparambil James, Marius Kloft, and Sophie Fellenz. 2024. Characterizing text datasets with psycholinguistic features. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14977–14990.

Mayank Kumar Nagda, Charu James, Marius Kloft, and Sophie Burkhardt. 2021. Hierarchical topic evaluation: Statistical vs. neural models. In *Bayesian Deep Learning Workshop at NeurIPS*.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.

Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy. 2021. TAN-NTM: Topic attention networks for neural topic modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3865–3880, Online. Association for Computational Linguistics.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2023. Antm: An aligned neural topic model for exploring evolving topics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. 2014. Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1977–1985.

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

Sandhya Subramani, Vaishnavi Sridhar, and Kaushal Shetty. 2018. A novel approach of neural topic modelling for document clustering. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2169–2173. IEEE.

Rohit Swami. 2020. All neurips (nips) papers.

Federico Tomasi, Mounia Lalmas, and Zhenwen Dai. 2022. Efficient inference for dynamic topic modeling with large vocabularies. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1950–1959. PMLR.

Luu Anh Tuan, Darsh Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9065–9072.

Chong Wang, David Blei, and David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456.

Minghao Wang and Paolo Mengoni. 2020. How pandemic spread in news: text analysis using topic model. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 764–770. IEEE.

Xiaobao Wu, Xinshuai Dong, Liangming Pan, Thong Nguyen, and Anh Tuan Luu. 2024a. Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3088–3105, Bangkok, Thailand. Association for Computational Linguistics.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024b. A survey on neural topic models: Methods, applications, and challenges. *ArXiv*, abs/2401.15351.

Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. *Advances in neural information processing systems*, 31.

Fang Yao and Yan Wang. 2020. Tracking urban geo-topics based on dynamic topic model. *Computers, Environment and Urban Systems*, 79:101419.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR.

Aonan Zhang, Jun Zhu, and Bo Zhang. 2013. Sparse online topic models. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1489–1500.

A Model Settings

This section describes the configuration of hyperparameters for each topic model under study. For

Dataset	NYT	UN	NIPS	NASA	Arxiv	DBLP
Domain	News	Politics	Science	Twitter	Science	Science
Number of Docs	273,938	273,398	276,657	9,703	129,984	151,232
Vocab Size	9,046	6,365	6,278	4,290	3,322	3,162
Timestamp	21	51	11	5	13	11
min_df	0.3%	0.05%	0.05%	0.05%	0.3%	0.3%
max_df	95%	95%	95%	95%	95%	95%

Table 2: Statistical analysis of corpus data. It demonstrates the datasets that vary in domain, number of documents, and vocabulary size. The NIPS is from (Swami, 2020), NYT is from (Sandhaus, 2008), UN is from (Jankin Mikhaylov et al., 2017b), Arxiv is from (arXiv.org submitters, 2023). Additionally, it displays the vocabulary size of documents that contain less than min_df percent of words and more than max_df percent of words.

the DETM (Dynamic Embedded Topic Model), word representations are derived using a skip-gram model with a 300-dimensional vector space. The DETM utilizes the perplexity score on the validation set as a criterion for termination. The learning rate for DETM is set at 0.001, and the hyperparameters delta, sigma, and gamma are fixed at 0.005, as recommended by the original authors. A uniform batch size of 100 is applied across all datasets. Additionally, the document corpus is segmented into training (80%), validation (10%), and testing (10%) subsets.

In the process of DLDA model training, the Gensim Python library's wrapper for Dynamic Topic Models (DTM) is utilized. Using this approach, all datasets are partitioned annually, thereby encapsulating each year's documents within a single temporal slice. For every dataset, the model undergoes 50 iterations, employing an alpha parameter set to 0.01. This alpha value is a critical hyperparameter in the Latent Dirichlet Allocation (LDA) models, influencing the degree of sparsity in the document-topic distributions across each time slice. Additionally, the top_chain_var parameter is set to 10. This setting plays a pivotal role in determining the variability in topic evolution over time within the DTM framework.

The default setting for Dynamic BERTopic was used, and the parameter evolution_tuning was set to true to display the topic evolution. Similarly, the default DNLDA model settings were used, tnd_iterations and lda_iterations being 500. The lda_beta value is 0.01, and the topic_depth value is 100. The results shown for all models are the average performance over three runs.

B Preprocessing Details

To preprocess the datasets (Monteiro et al., 2024), the text was converted to lowercase, and both stopwords and punctuation were removed. As part of tokenization, we used Spacy (Honnibal and Montani, 2017). For removing words that are present in fewer than or more than min/max percent of documents, we used Scikit-learn (Kramer and Kramer, 2016). The UN and NIPS corpus has been broken down into paragraphs, with each paragraph being treated as a separate document. The statistical analysis of the document corpus is shown in table 2.

C Additional Qualitative Assessment on NYT Dataset

To compare how CoNTM, Dynamic BERTopics, DETM (Dynamic Embedded Topic Model), and DLDA (Dynamic Latent Dirichlet Allocation) capture emerging topics with good topic coherence, we should look at the clarity, continuity, and thematic relevance of the topics they produce over time. In the main paper section 4.6, the CoNTM and Dynamic BERTopics are explained.

Evolving Topics: As shown in Figure 6, DETM's topics vary over the years, focusing on the article, king, page, and lead in 1990. The topic shifted to the election of Clinton in 1992. In DLDA, we can also observe topic changes over the years. For example, in 1990, the topics were centered around "republican", "campaign" and "political" while by 1992, the focus shifted to "clinton" "campaign", "president" and "republican". Even though both models shift the topic to the election of Clinton in 1992, the topic coherence of DLDA is better than DETM for this topic.

Topic Coherence: The DETM demonstrates good topic coherence, with closely related terms consistently appearing across the years. The topics reflect a stable and focused exploration of the topic politics, with less dramatic shifts compared to Dynamic BERTopics. Furthermore, the DLDA topics are less coherent, with each year building

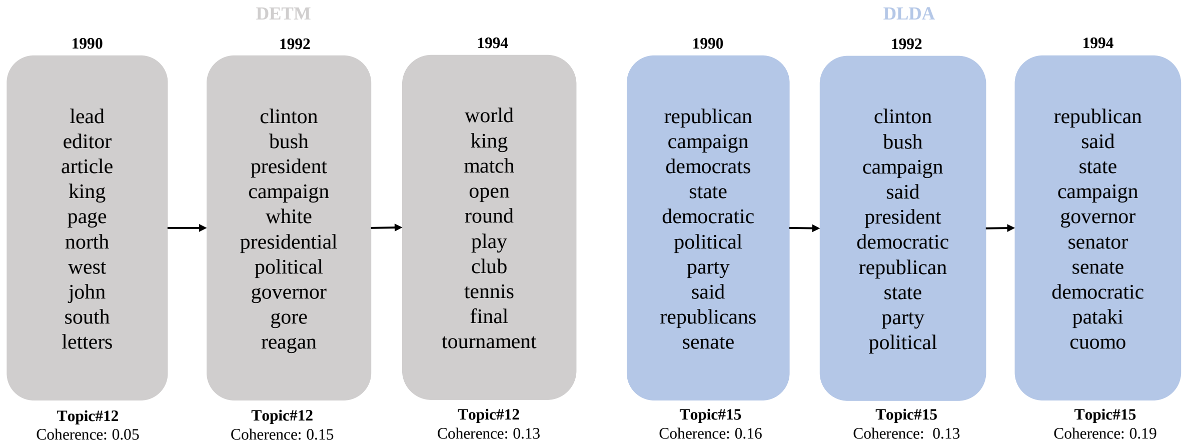


Figure 6: The figure illustrates the progression of topics#12 (DETM) and topic#15 (DLDA) in the NYT dataset, spanning the years 1990 to 1994, both of which focus on politics (election of Clinton in 1992). The DLDA models reflect stable and clear thematic focus topics.

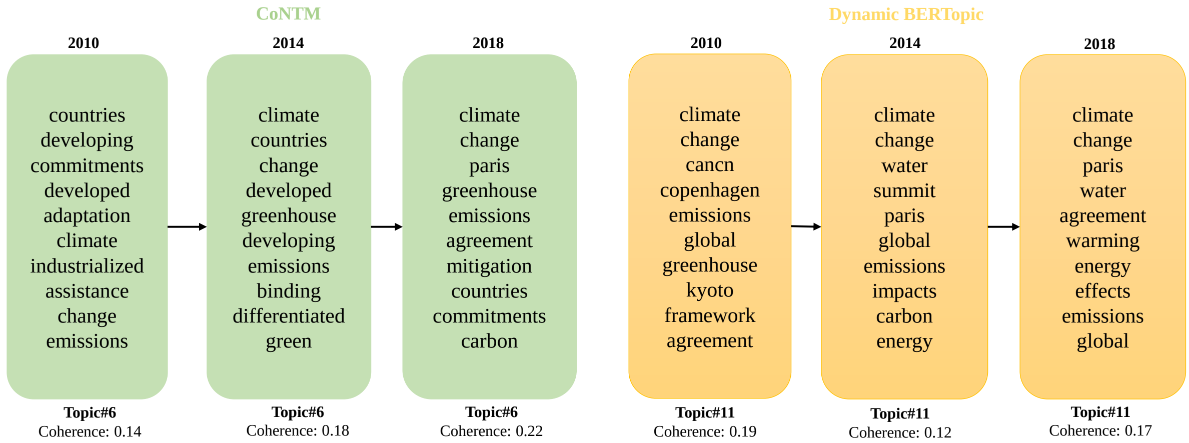


Figure 7: The figure illustrates the progression of topics#6 (CoNTM) and topic#11 (Dynamic BERTopics) in the UN dataset, spanning the years 2010 to 2018, both of which focus on climate change. The CoNTM model captures the emerging topic, transitioning from climate change in developed countries to the Paris Agreement which was adopted in 2015 during the UN climate change conference.

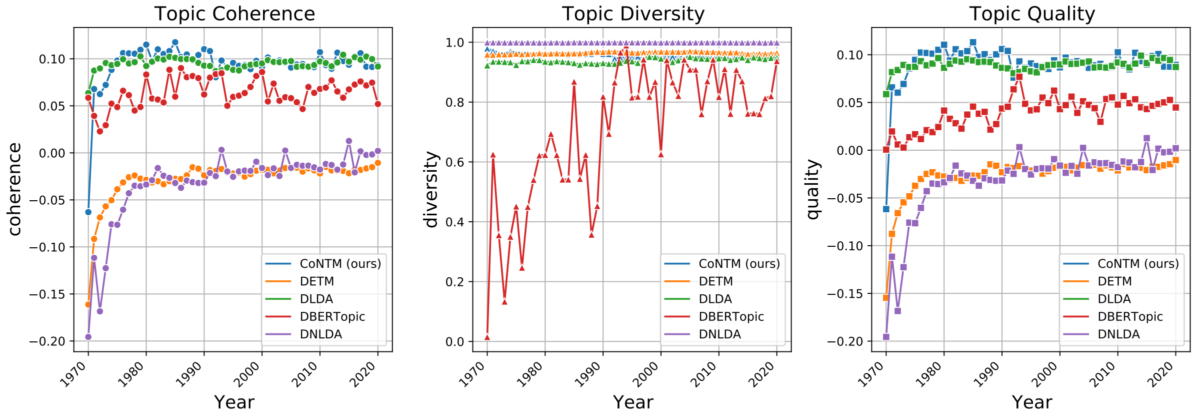


Figure 8: The figure illustrates how three measures—coherence, diversity, and topic quality—change yearly for the UN Dataset. In the case of the CoNTM model, coherence improves over time and remains stable, similar to the DLDA model. On the other hand, the topic quality of the Dynamic BERT model also increases over time, but it is lower compared to both the CoNTM and DLDA models.

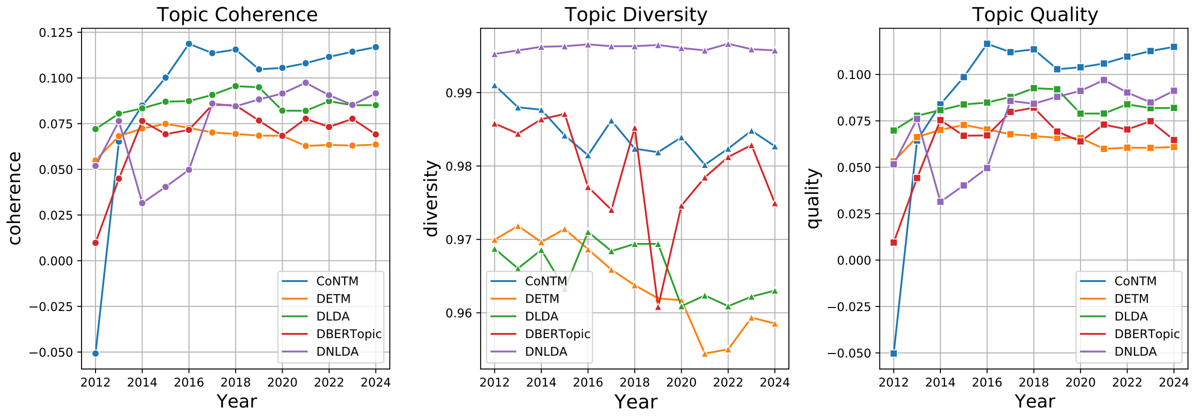


Figure 9: The figure illustrates how three measures—coherence, diversity, and topic quality—change yearly for the Arxiv Dataset. The topic quality of the CoNTM model improves over time and outperforms other models in terms of overall topic quality.

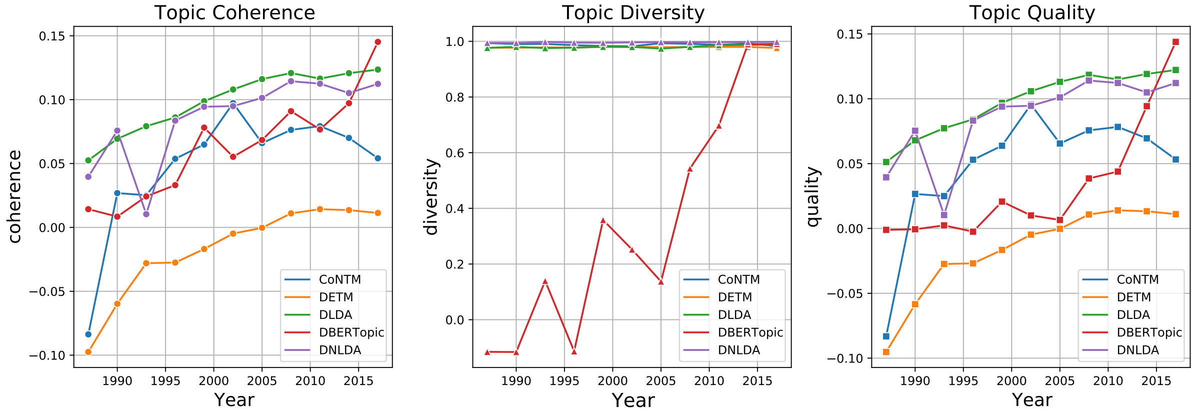


Figure 10: The figure illustrates how three measures—coherence, diversity, and topic quality—change yearly for the NIPS Dataset. The CoNTM model exhibits moderate topic quality compared to other models; this is due to insufficient documents available at the early timestamp.

on the previous one, showing gradual changes in politics. DLDA maintains a clear thematic focus, although the evolution of topics is less coherent than in CoNTM.

In summary, while all models effectively capture topic evolution, CoNTM and DETM provide more coherent and gradual thematic progressions, the DLDA shows stability in topic focus, and Dynamic BERTopics excels in capturing broader shifts. CoNTM, in particular, demonstrates a strong capability for capturing emerging topics with good coherence, making it particularly useful for understanding detailed shifts in a specific thematic area over time. Example emerging topics are shown in Figure 19 for the NYT dataset.

D Qualitative Assessment on UNDebates Dataset

To compare the CoNTM model with other models, we examine evolving topics and topic coherence using the UN dataset. Figure 7 shows how the topic (climate change) evolves over time for the CoNTM and Dynamic BERTopic models. In both models, we can see the topic shift to the Paris Agreement in 2018, and overall topic coherence remains the same for this topic. Figure 20 shows three emerging topics on climate change, war, and human rights for the CoNTM model. Here, the x-axis is the timestamps, and the y-axis is the word probability. Regarding climate change, the word probability of pairs, agreement, emission, and greenhouse increased dramatically towards 2015. This is due to the Paris Agreement adopted by the UN for climate change. As shown in Figure 20, the topic "war" has a high word probability for the words conflict, Iran, and Iraq, indicating a conflict in 1986.

Furthermore, Figure 11 illustrates the evolution of the climate change topic in the UN dataset for the DETM and DLDA models. In both models, the topic shifts to the Paris Agreement in 2018.

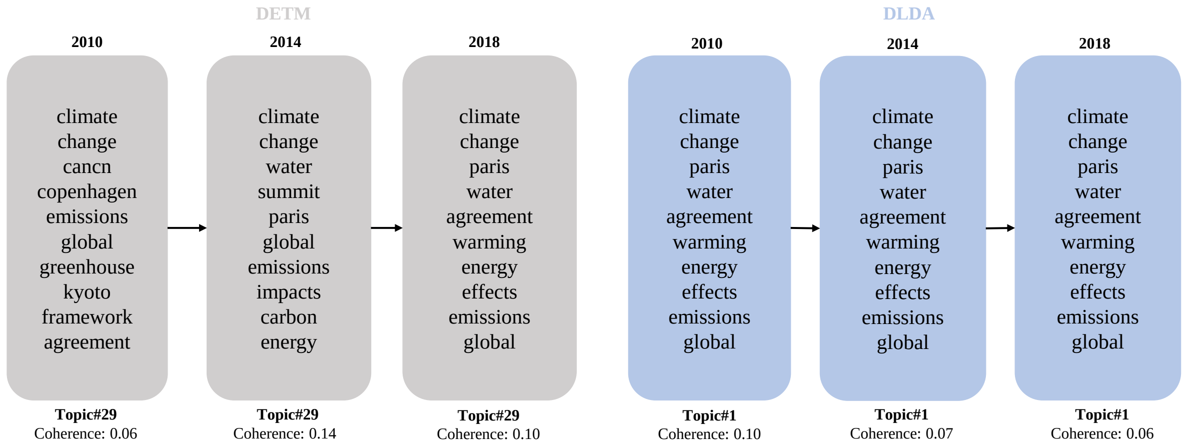

DETM
2010
2014
2018
climate change cancn copenhagen emissions global greenhouse kyoto framework agreement
climate change water summit paris global emissions impacts carbon energy
climate change paris water agreement warming energy effects emissions global
Topic#29
Coherence: 0.06
Topic#29
Coherence: 0.14
Topic#29
Coherence: 0.10
DLDA
2010
2014
2018
climate change paris water agreement warming energy effects emissions global
climate change paris water agreement warming energy effects emissions global
climate change paris water agreement warming energy effects emissions global
Topic#1
Coherence: 0.10
Topic#1
Coherence: 0.07
Topic#1
Coherence: 0.06

Figure 11: The figure illustrates the progression of topics#29 (DETM) and topic#1 (DLDA) in the UN dataset, spanning the years 2010 to 2018. Both topics focus on climate change, with the DETM model showing better topic coherence for this subject compared to the DLDA model.

Additionally, the DETM model's topic coherence is slightly better than the DLDA model for this topic.

In conclusion, the CoNTM model seems to produce more specific topics, while the Dynamic BERTopics generates broader thematic topics, indicating a difference in the level of transition each model offers. Also, the CoNTM captures the emerging topic, shifting from climate change in developed countries to the Paris Agreement in 2018, with an overall TTS score of 0.49 (see table 6) for the UN dataset. In contrast, in Dynamic BERTopics, the temporal shift is more (TTS is 0.31) with a lower coherence score. The DLDA model (Figure 11) shows a low coherence score for the climate change topic. Additionally, the TTS is 0.64 (see Table 6), which means that the topics are changing more smoothly. A higher TTS value indicates that the topic is not evolving.

E Stability Across Varying Topic Size

The results in Table 3 demonstrate that CoNTM exhibits remarkable stability across varying topic sizes and datasets, consistently maintaining top or near-top ranks in all three evaluation metrics (TC, TD, and TQ). Its average ranks of 1.50 (TC), 1.83 (TD), and 1.33 (TQ) indicate not only strong absolute performance but also low sensitivity to the number of topics, implying that CoNTM can preserve topic coherence and quality as topic granularity changes.

In contrast, baseline models show more fluctuation in performance. DETM and DLDA deliver moderate but less competitive results, indicating consistent yet less stable behavior. CFDTM performs strongly in TC and TQ but shows greater variability in TD, suggesting some instability in preserving topic distinctness. DNLDA achieves the best diversity score but fails to maintain coherence and overall topic quality. DBERTopic exhibits the weakest stability, with relatively high variance across metrics and datasets. Overall, CoNTM offers the most stable and balanced performance, effectively preserving both topic coherence and diversity across different topic sizes, which is crucial for robust temporal topic modeling.

F Sensitivity Analysis of Rho (ρ)

We conduct a sensitivity analysis on rho ($\rho_t = \frac{1}{(\tau_0+t)^\kappa}$) from Algorithm 2, step 10. Figure 13 shows that changes in the parameter ρ with respect to κ do not severely impact performance. The y-axis represents the varying κ value, which affects the corresponding ρ values. The x-axis represents different timestamps. The figure shows the effect of ρ on predictive perplexity on future timestamps and topic quality over time. On average, the results indicate that ρ has minimal impact on predictive perplexity or topic quality. The effect of κ and τ for other datasets is shown in Figure 12, and Figure 14, showing that a κ value of 0.7 and a τ value of 1 are optimal.

Figure 12 illustrates that variations in ρ with respect to τ have a minimal impact on performance. The y-axis represents the varying τ value, which influences the corresponding ρ values. The x-axis represents different timestamps. The figure shows the effect of τ on predictive perplexity at future

	CoNTM (ours)			DETM			DLDA			DBERTopic			DNLDA			CFDTM		
Dataset	TC ↑	TD ↑	TQ ↑	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ
NIPS	.095	.996	.095	.082	.986	.081	.085	.951	.082	.068	.759	.018	-.099	.979	-.089	.123	.978	.120
NYT	.181	.996	.181	.118	.980	.115	.098	.909	.088	.091	.796	.071	-.077	1.00	-.077	.179	.996	.178
UN	.116	.975	.116	.094	.972	.092	.090	.899	.081	.055	.727	.032	-.053	.999	-.053	.123	.951	.116
Tweets	.094	.978	.090	.054	.971	.052	.084	.931	.079	.101	.967	.094	-.162	.996	-.162	.087	.968	.084
Arxiv	.113	.976	.111	.069	.965	.067	.059	.916	.054	.054	.927	.019	-.090	1.00	-.090	.102	.953	.097
DBLP	.126	.973	.123	.080	.955	.076	.074	.907	.068	.087	.938	.031	-.080	1.00	-.080	.115	.927	.106
Av. Rank	**1.50**	1.83	**1.33**	3.50	3.50	3.50	3.83	5.50	3.83	4.33	4.83	4.33	6.00	**1.33**	6.00	1.83	3.83	2.00

Table 3: The table compares the performance of six topic modeling algorithms (20 topics): CoNTM, DETM, DLDA, Dynamic Bertopic, DNLDA, and CFDTM based on topic coherence (TC), topic diversity (TD), and topic quality (TQ) with the temporal reference corpus. These evaluations were conducted across six diverse datasets: NIPS, New York Times, UN, NASA Tweets, and Arxiv.

timestamps and topic quality over time. The results indicate that a τ value of 1 yields lower predictive perplexity and higher topic quality for the CoNTM model.

Figure 14 shows the sensitivity analysis of the parameter ρ with respect to κ for the NIPS, Arxiv, and Tweets datasets. The results indicate that κ does not significantly impact topic quality and predictive perplexity. However, on average, a κ value of 0.7 yields better results.

G Statistical Significance Testing

The t-test results indicate significant differences in perplexity values for some datasets, while others do not show substantial differences. For datasets such as NIPS, NYT, and DBLP, the p-values are well below 0.05, indicating a statistically significant difference between the models. However, for datasets like un and tweets, the p-values are above 0.05, suggesting no strong evidence for a difference.

Dataset	t-statistic	p-value
NIPS	-8.716	6.365591e-08
NYT	-21.555	1.828662e-13
UN	-4.014	5.186161e-04
Tweets	-0.928	3.594985e-01
Arxiv	0.829	4.128483e-01
DBLP	4.170	1.955962e-04

Table 4: Statistical Significance Test (t-test) on predictive perplexity between CoNTM and DETM with twenty different seed values.

H Runtime Comparison

A comparative analysis of their runtimes is essential in evaluating the efficiency of various topic modeling algorithms (see Figure 15). This section discusses the runtime performance of four different models, CoNTM (our proposed model), DETM, DLDA, DBERTopic, and CFDTM, across two datasets: the UN Dataset and the Arxiv Dataset.

For the UN Dataset, the runtime analysis reveals a significant variance in the computational efficiency of the models. The CoNTM model shows a remarkable performance advantage, with the lowest runtime among the models tested. Following CoNTM, DBERTopic, CFDTM, DLDA, and DETM exhibit moderately higher runtimes.

The runtime comparison maintains a similar trend in the context of the Arxiv dataset. The CoNTM model again stands out for its efficiency, underlining the effectiveness of our optimization strategies in reducing computational overhead. The other models, DBERTopic, DETM, and DLDA, follow in increasing order of runtime, consistent with the findings from the UN Datasets.

I Topic Coherence vs Diversity

The quantitative analysis of additional datasets such as NIPS, UN, and Tweets is shown in Figure 18. On the Tweets dataset, DBERTopic demonstrates notably high topic quality. This can be due to its use of pre-trained word embeddings, which significantly enhance topic coherence, especially in smaller datasets. The use of pre-trained embeddings allows DBERTopic to capture semantic relationships more effectively, leading to improved

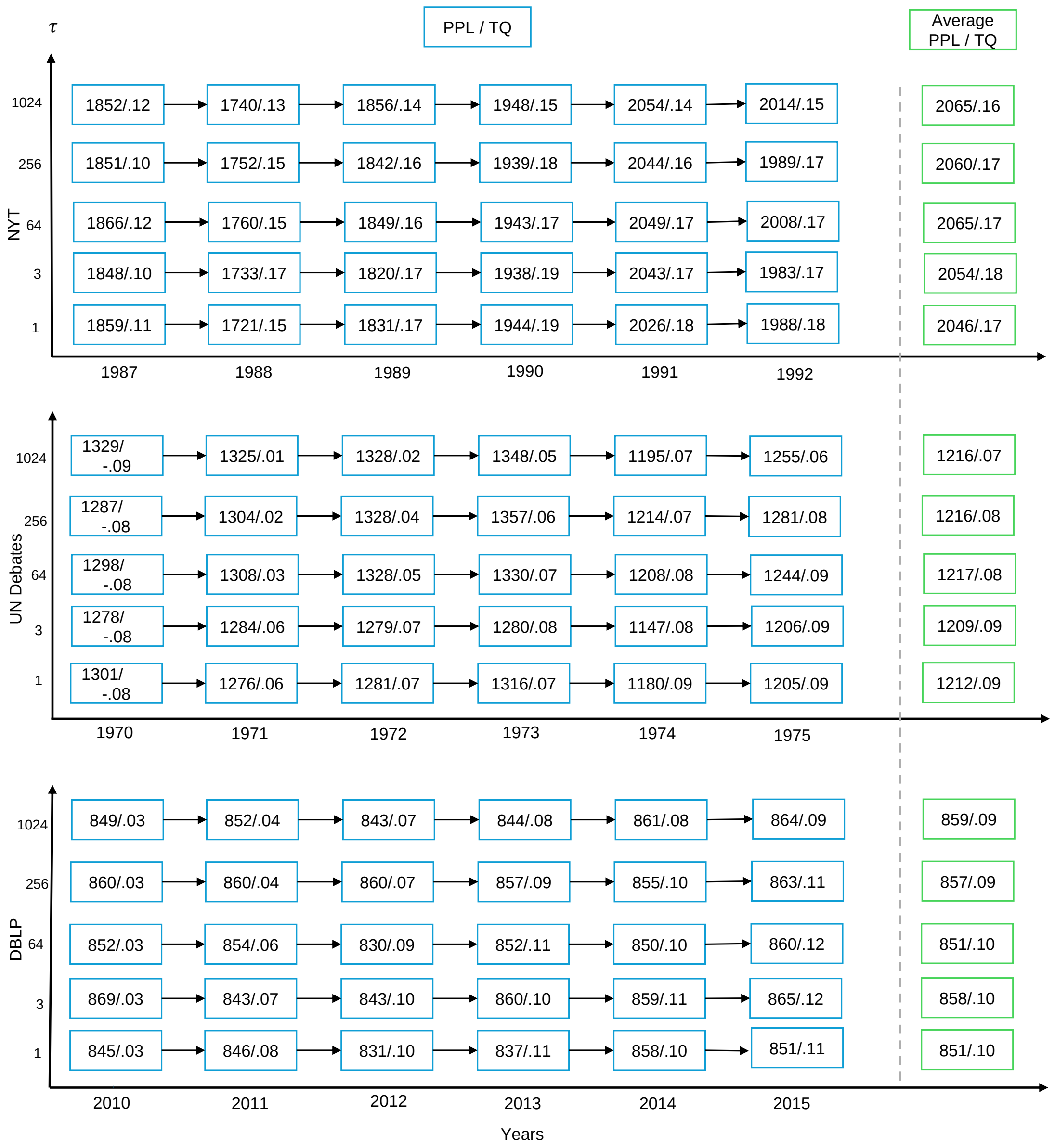


Figure 12: The figure shows the sensitivity analysis on the parameter ρ ($\rho_t = \frac{1}{(\tau_0+t)^\kappa}$) from Algorithm 2, demonstrating a robustness to the exact choice of ρ. The x-axis shows perplexity/topic quality (PPL/TQ) changes over time for three different datasets such as DBLP, UnDebates, and NYT. The y-axis represents varying τ values. Here, perplexity (PPL) is the predictive perplexity on future timestamps. The blue box shows the PPL/TQ score for five timestamps, and the green box shows the average PPL/TQ for all timestamps across the respective datasets.

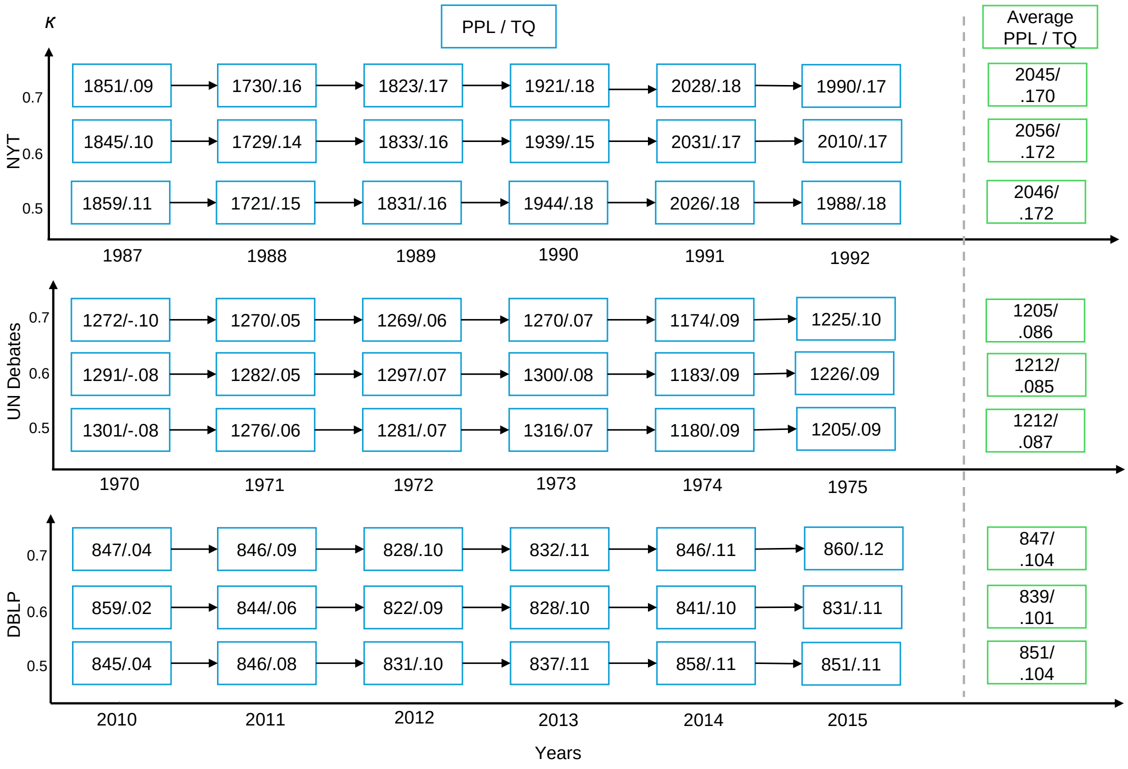
performance in terms of topic quality and coherence compared to other models on limited data.

J Additional Quantitative Results

This section analyzes CoNTM and compares it with other models on all datasets.

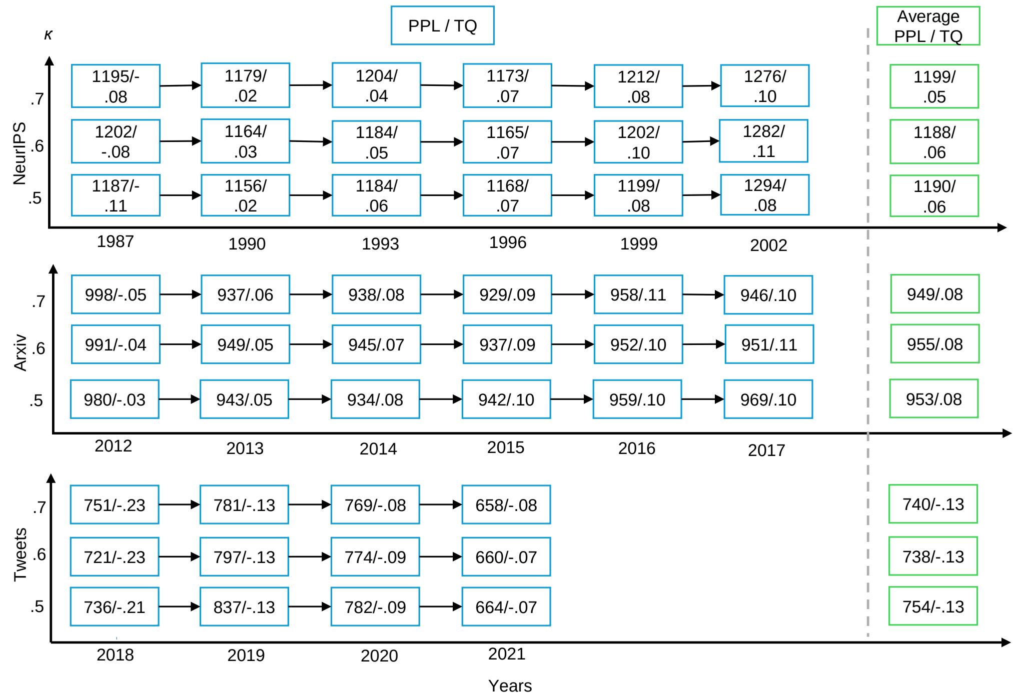
Topic Quality vs Predictive Perplexity: In the NeurIPS, Tweets, and DBLP results, a clear pattern emerges (see Figure 17). CoNTM consistently offers the strongest balance between topic quality and predictive perplexity, with CFDTM performing almost as well and in some cases matching it. In contrast, DETM often achieves lower perplexity but fails to reach the same levels of topic quality, while DLDA shows more average performance, never surpassing the leading models. On Tweets, DBERTopic achieves good topic quality, comparable to the stronger models, but this comes at the cost of higher perplexity.

The NYT, UN Debates, and Arxiv datasets reveal a similar tendency (see Figure 16). CoNTM and CFDTM again dominate the tradeoff space, managing to combine higher topic quality with competitive perplexity values, which positions them



Dataset	κ							Average PPL / TQ
NYT	0.7	1851/.09	1730/.16	1823/.17	1921/.18	2028/.18	1990/.17	2045/ .170
	0.6	1845/.10	1729/.14	1833/.16	1939/.15	2031/.17	2010/.17	2056/ .172
	0.5	1859/.11	1721/.15	1831/.16	1944/.18	2026/.18	1988/.18	2046/ .172
		1987	1988	1989	1990	1991	1992	
UN Debates	0.7	1272/-.10	1270/.05	1269/.06	1270/.07	1174/.09	1225/.10	1205/ .086
	0.6	1291/-.08	1282/.05	1297/.07	1300/.08	1183/.09	1226/.09	1212/ .085
	0.5	1301/-.08	1276/.06	1281/.07	1316/.07	1180/.09	1205/.09	1212/ .087
		1970	1971	1972	1973	1974	1975	
DBLP	0.7	847/.04	846/.09	828/.10	832/.11	846/.11	860/.12	847/ .104
	0.6	859/.02	844/.06	822/.09	828/.10	841/.10	831/.11	839/ .101
	0.5	845/.04	846/.08	831/.10	837/.11	858/.11	851/.11	851/ .104
		2010	2011	2012	2013	2014	2015	

Figure 13: The figure shows the sensitivity analysis on the parameter ρ ($\rho_t = \frac{1}{(\tau_0+t)^\kappa}$) from Algorithm 2, demonstrating a robustness to the exact choice of ρ. The blue box shows the PPL/TQ score for five timestamps, and the green box shows the average PPL/TQ for all timestamps across the respective datasets. It shows that ρ does not have a significant impact on PPL/TQ.



Dataset	κ							Average PPL / TQ
NeurIPS	.7	1195/- .08	1179/ .02	1204/ .04	1173/ .07	1212/ .08	1276/ .10	1199/ .05
	.6	1202/ -.08	1164/ .03	1184/ .05	1165/ .07	1202/ .10	1282/ .11	1188/ .06
	.5	1187/- .11	1156/ .02	1184/ .06	1168/ .07	1199/ .08	1294/ .08	1190/ .06
		1987	1990	1993	1996	1999	2002	
Arxiv	.7	998/-.05	937/.06	938/.08	929/.09	958/.11	946/.10	949/.08
	.6	991/-.04	949/.05	945/.07	937/.09	952/.10	951/.11	955/.08
	.5	980/-.03	943/.05	934/.08	942/.10	959/.10	969/.10	953/.08
		2012	2013	2014	2015	2016	2017	
Tweets	.7	751/-.23	781/-.13	769/-.08	658/-.08			740/-.13
	.6	721/-.23	797/-.13	774/-.09	660/-.07			738/-.13
	.5	736/-.21	837/-.13	782/-.09	664/-.07			754/-.13
		2018	2019	2020	2021			

Figure 14: The figure shows the sensitivity analysis on the parameter ρ ($\rho_t = \frac{1}{(\tau_0+t)^\kappa}$) from Algorithm 2, demonstrating a robustness to the exact choice of ρ. The x-axis shows perplexity/topic quality (PPL/TQ) changes over time for three different datasets such as NIPS, Arxiv, and Tweets. The y-axis represents varying κ values. Here the perplexity (PPL), is the predictive perplexity on future timestamp. The blue box shows the PPL/TQ score for five timestamps, and the green box shows the average PPL/TQ for all timestamps across the respective datasets.

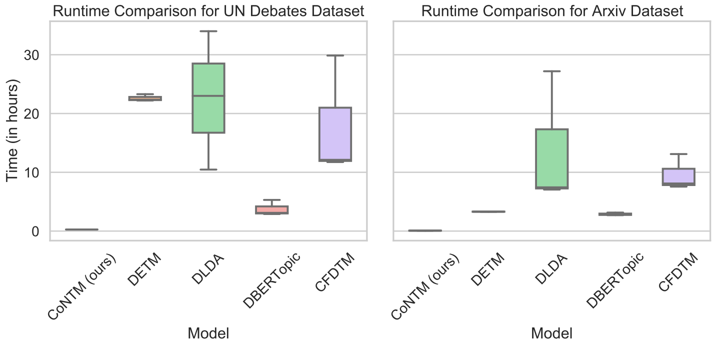


Figure 15: Runtime comparison (in hours) for evolving topics on UN and the Arxiv dataset shows that the CoNTM model has the shortest runtime, outperforming all other tested models.

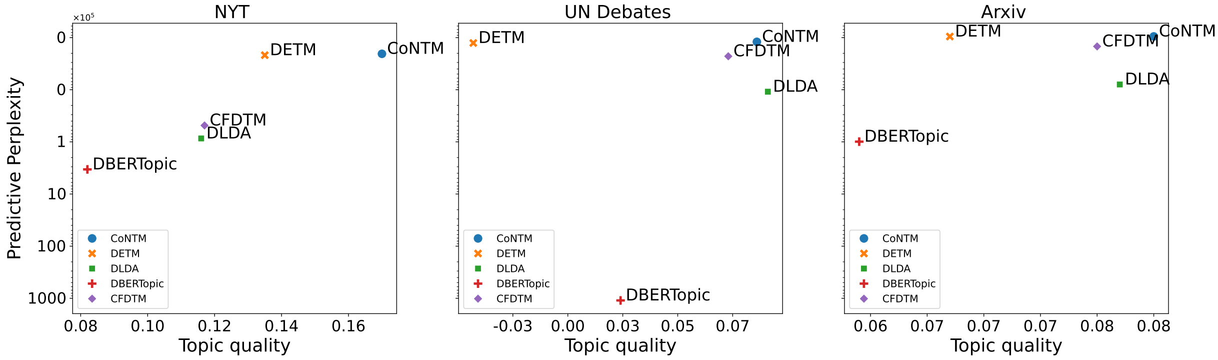


Figure 16: The figure shows the model performance quantitatively for NYT, UN Debates, and Arxiv datasets. In the plot, the top-right corner indicates that the model achieves high topic quality and low predictive perplexity. Our model (CoNTM) outperformed the other models in terms of both quality and perplexity.

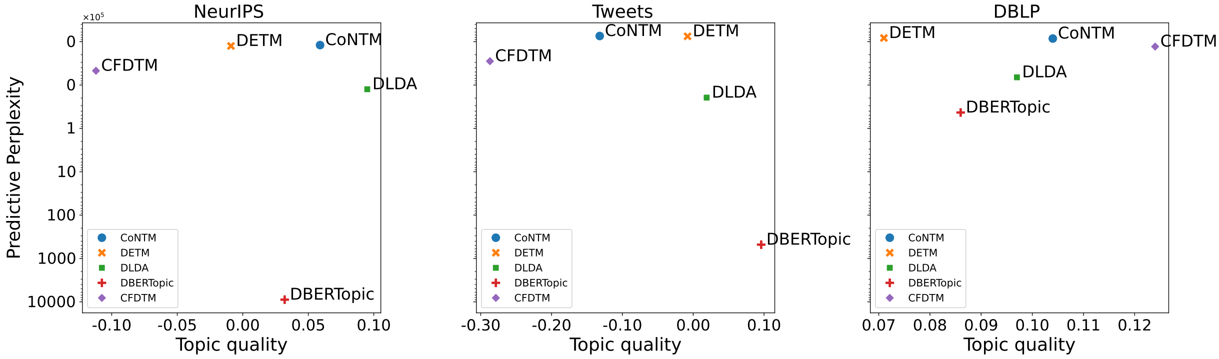


Figure 17: Comparison of topic quality versus predictive perplexity (logarithmic scale) across NeurIPS, Tweets, and DBLP datasets. CoNTM and CFDTM consistently achieve a superior balance of higher topic quality and lower perplexity, while DBERTopic shows weak performance across all datasets.

as the most effective approaches. DETM remains somewhat reliable in perplexity but trails in quality, and DLDA performs moderately without excelling. DBERTopic once more lags far behind, demonstrating weak results across the board.

K Emerging Topics

This section provides a few emerging topics in the NYT and UN dataset. Figure 19 shows the emerging topic "politics," with word probability on the y-axis and timestamps on the x-axis for the NYT dataset. In 1997, *Clinton* was inaugurated for his second term as the 42nd President of the U.S., with his final years in office spanning from 1999 to 2000. Following this period, there is a noticeable decline in the word probability of *Clinton*.

Figure 20 shows the emerging topics "Climate Change", "War", and "Human Rights" from the UN dataset. For the "Climate Change" topic, the

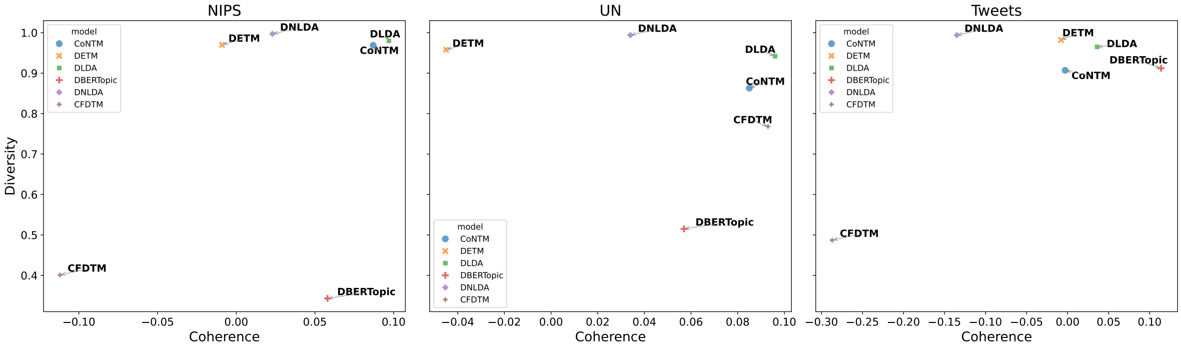


Figure 18: The figure shows the model performance quantitatively for the NIPS, UN, and Tweets datasets. The top-right corner indicates that the model achieves high topic coherence and diversity.

word probability of Paris, agreement, emissions, and greenhouse is notably high, aligning with the negotiations and implementation of the Paris Agreement in 2015.

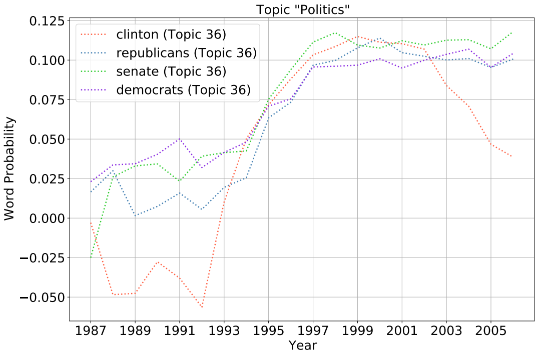


Figure 19: The figure shows emerging topics "Politics" in the CoNTM (our) model for the NYT dataset. The y-axis shows the word probability of topic words from the trained model and the x-axis shows the timestamp.

L Generalization via Predictive Perplexity (PPL)

Perplexity is a standard metric for evaluating probabilistic language models. Table 5 shows predictive perplexity on future timestamp data for all models, where we compute the perplexity of documents in timestamp $(t+1)$ based on the model trained at timestamp (t). The CoNTM model achieves the lowest predictive perplexity (lower is better) in all datasets except Tweets, demonstrating its ability to predict unseen data. The Tweets are very short, providing limited context for the model to capture meaningful patterns, leading to a decrease in performance compared to longer documents. In conclusion, CoNTM demonstrates good performance, while DLDA and Dynamic BERTopic show significantly lower performance. The t-test for CoNTM and DETM is detailed in Appendix G. Also, the value presented in the table is the average of three random seeds. The DNLDA perplexity score is not included in the table due to its poor performance. Figures 23, and 24 show the trade-off between predictive perplexity and topic quality on the Arxiv and DBLP datasets. The results shown are based on three randomly selected seeds. In the figure, when $\alpha = 0.90$, topic quality (where higher is better) is high, but predictive perplexity (where lower is better) is also high. Conversely, when $\alpha = 0.10$, topic quality decreases, but predictive perplexity improves. This illustrates a trade-off between topic quality and predictive perplexity.

Dataset	CoNTM	DETM	DLDA	DBERTopic	CFDTM
NIPS	**1199**	1254	12.5K	8.9E8	4686
NYT	**2045**	2172	85.9K	337K	48.4K
UN	**1205**	1268	10.9K	1.1E8	2279
Tweets	**740**	752	19.5K	4.8E7	2821
Arxiv	**949**	953	7921	99K	1475
DBLP	847	**823**	6634	43K	1297

Table 5: The table shows the average predictive perplexity over three runs for six diverse datasets. The CoNTM model provides better predictive perplexity on almost all datasets.

M Temporal Quality and Smoothness

An analysis of the NYT dataset from 1987 to 2007 was conducted to determine the temporal quality of the dataset. Each model's coherence, diversity, and overall quality were tracked over the years (see Figure 3). Our model (CoNTM) maintained stable performance across the years, with an increase in coherence and quality towards the later years. As a result of incorporating continual learning into the model, the quality of topics has increased over the years, in contrast to other models. DBERTopic and DNLDA exhibit fluctuations in performance. While topic coherence for our model increases with time, it decreases for Dynamic BERTopic.

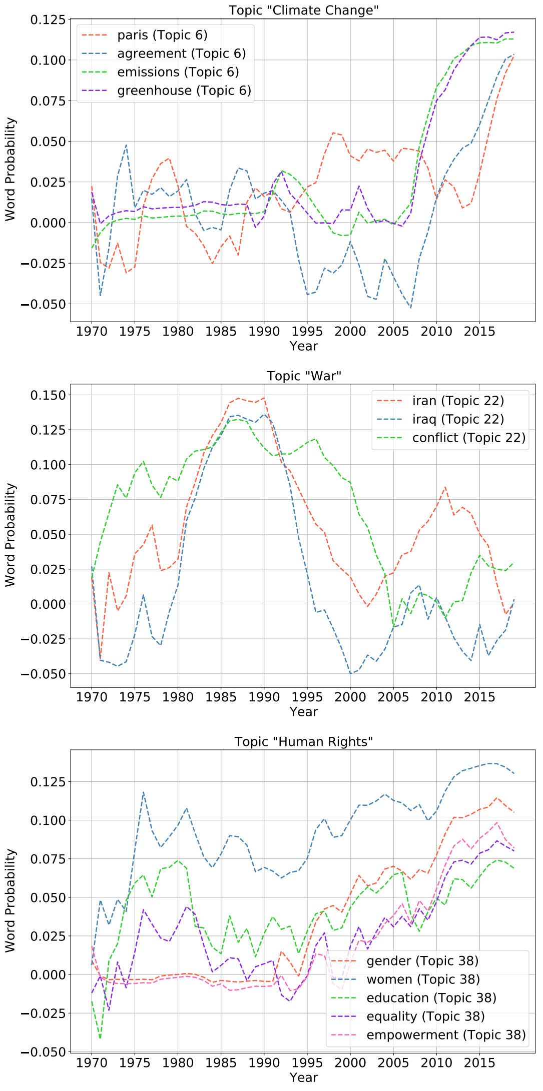


Figure 20: The figure shows three emerging topics "Climate Change", "War", and "Human Rights" in the CoNTM (our) model for the UN dataset. The y-axis shows the word probability of topic words from the trained model and the x-axis shows the timestamp.

Additionally, Dynamic BERTopic shows a drastic change in topic diversity, which negatively impacts topic quality. The DETM and DLDA show stable diversity and quality over time. Additionally, the DNLDA model shows an improvement in topic quality, though it remains significantly lower than the others. Furthermore, the temporal quality of UN from 1970 to 2020 can be seen in Figure 8. The temporal characteristics of the Arxiv dataset are illustrated in Figure 9. Similarly, the temporal characteristics of the NIPS dataset are shown in Figure 10 in Appendix D.

For each dataset, the score for temporal topic smoothness (TTS) can be found in Table 6. On average, the CoNTM model has a TTS score of 0.49, with an exception on the Tweets dataset. The Tweets have a low TTS score because the dataset lacks sufficient documents to learn more coherent and diverse topics. In summary, while the topics change gradually, the transitions are not completely smooth, allowing us to observe the evolution of topics. This is because our model learns new topics at each timestamp without forgetting previously learned information.

Models	NIPS	NYT	UN	Tweets	Arxiv	DBLP
DETM	.802	.646	.870	.820	.864	.866
DLDA	.602	.713	.635	.372	.655	.748
DBERTopic	.272	.360	.307	.186	**.495**	**.547**
DNLDA	.016	.008	.019	.017	.013	.015
CFDTM	.753	.892	.902	.897	.969	.969
CoNTM (ours)	.368	**.569**	**.496**	.191	**.498**	**.599**

Table 6: The tables display Temporal Topic Smoothness (TTS) scores for four models across various datasets. Our model averages a TTS score of 0.51, except for the Tweets dataset, indicating that its topic transitions are balanced. The bolded number indicates the TTS score closest to 0.50 ± 10.

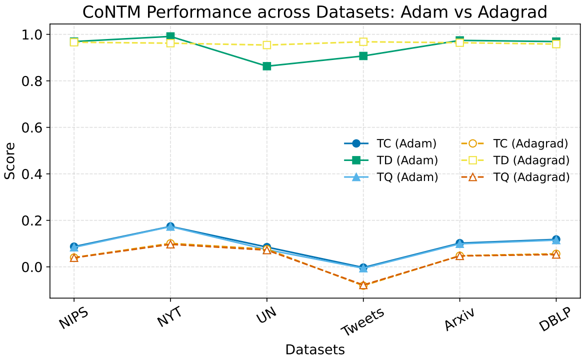


Figure 21: Comparison of CoNTM performance with Adam and Adagrad optimizers across six datasets. Metrics include Topic Coherence (TC), Topic Diversity (TD), and Topic Quality (TQ). Adam yields higher TC and TQ, indicating more coherent topics, while both optimizers show similar diversity.

N Optimizer and Learning Rate Sensitivity Analysis

This section investigates the impact of optimization strategies and learning rate configurations on the performance of the Continuous Neural Topic Model (CoNTM). We compare the Adam and Adagrad optimizers (see Figure 21) across six benchmark datasets such as NIPS, NYT, UN, Tweets, Arxiv, and DBLP, evaluating topic coherence (TC), topic diversity (TD), and topic quality (TQ). The results show that Adam consistently achieves higher TC and TQ values, indicating more coherent and semantically meaningful topics, while both optimizers maintain comparable topic diversity.

We further analyze CoNTM's sensitivity to learning rates (Figure 22), showing how TC, TD, and TQ vary across datasets and emphasizing the importance of proper optimizer and learning rate selection for stable, high-quality topics.

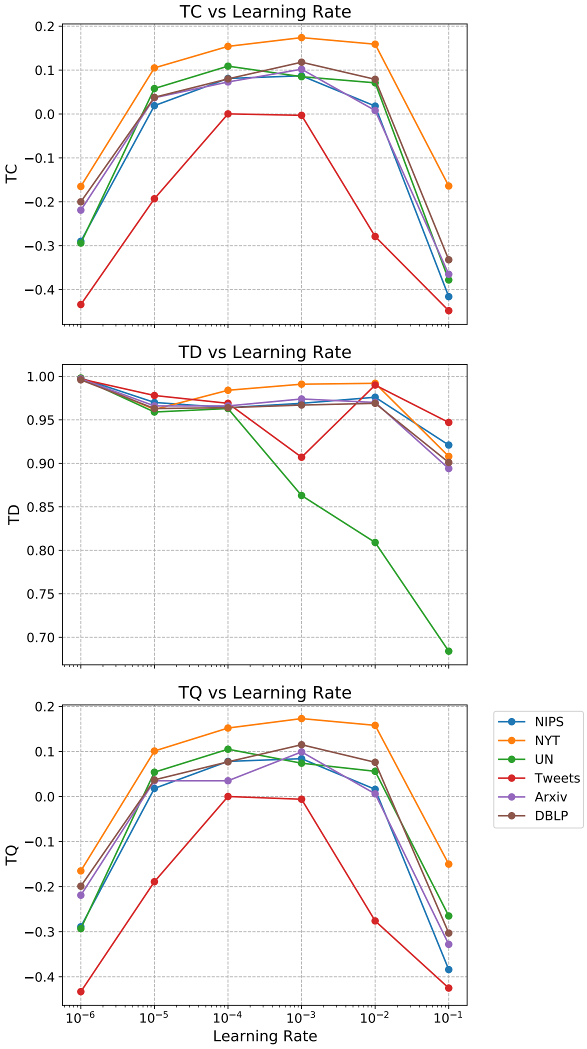


Figure 22: Performance of topic coherence (TC), topic diversity (TD), and topic quality (TQ) across different learning rates on multiple datasets (NIPS, NYT, UN, Tweets, Arxiv, DBLP).

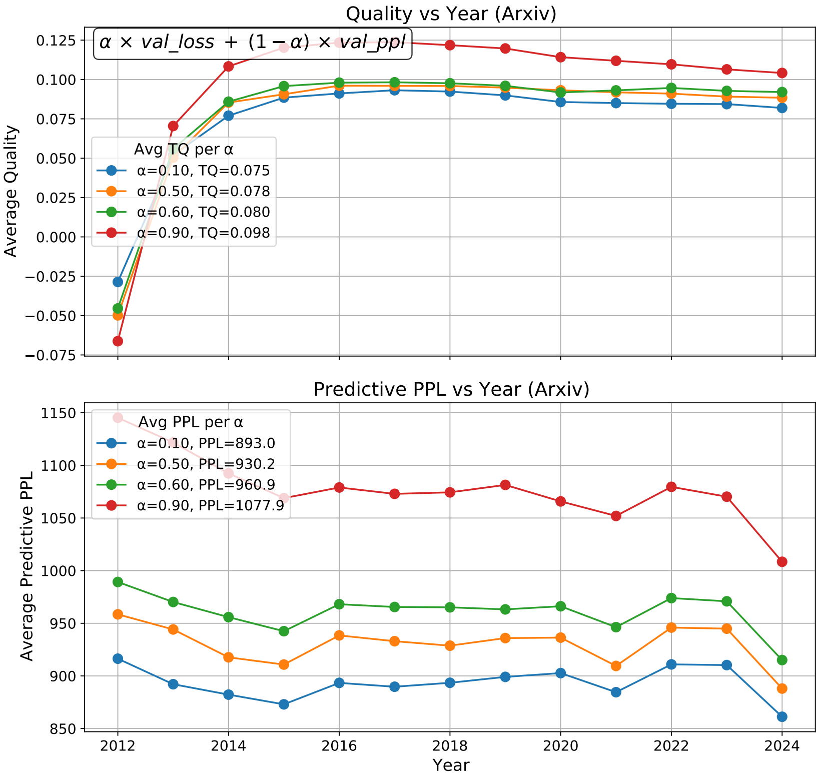


Figure 23: The figure illustrates the trade-off between predictive perplexity and topic quality for Arxiv dataset.

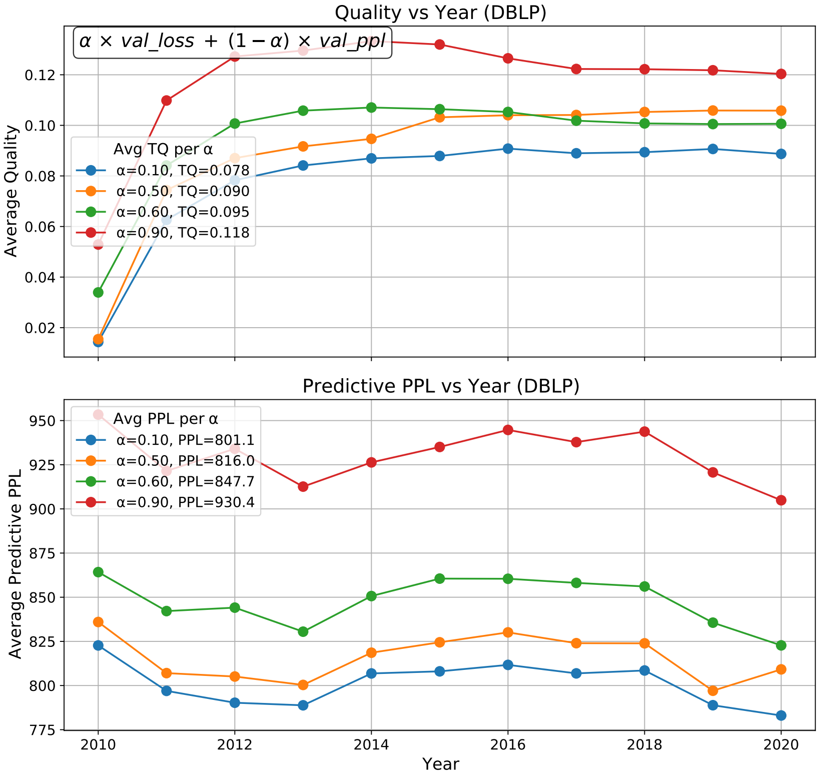


Figure 24: The figure illustrates the trade-off between predictive perplexity and topic quality for DBLP dataset.