Towards the TOPMOST: A Topic Modeling System Toolkit

Xiaobao Wu  Fengjun Pan  Anh Tuan Luu
Nanyang Technological University
xiaobao002@e.ntu.edu.sg  fengjun001@e.ntu.edu.sg  anhtuan.luu@ntu.edu.sg

Abstract

Topic models have a rich history with various applications and have recently been reinvigorated by neural topic modeling. However, these numerous topic models adopt totally distinct datasets, implementations, and evaluations. This impedes quick utilization and fair comparisons, and thereby hinders their research progress and applications. To tackle this challenge, we in this paper propose a Topic Modeling System Toolkit (TOPMOST). Compared to existing toolkits, TOPMOST stands out by supporting more extensive features. It covers a broader spectrum of topic modeling scenarios with their complete lifecycles, including datasets, preprocessing, models, training, and evaluations. Thanks to its highly cohesive and decoupled modular design, TOPMOST enables rapid utilization, fair comparisons, and flexible extensions of diverse cutting-edge topic models. These improvements position TOPMOST as a valuable resource to accelerate the research and applications of topic models. Our code, tutorials, and documentation are available at https://github.com/bobxwu/topmost. Our demo video is at https://youtu.be/9bN-rs4Gu3E?si=LunquCRhBZwyd1Xg.

<table>
<thead>
<tr>
<th>Topic Modeling Scenario</th>
<th>OCTIS TOPMOST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic topic modeling</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>Datasets ✓ Models ✓ Evaluations ✓</td>
<td></td>
</tr>
<tr>
<td>Hierarchical topic modeling</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>Datasets ✓ Models ✓ Evaluations ✓</td>
<td></td>
</tr>
<tr>
<td>Dynamic topic modeling</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>Datasets ✓ Models ✓ Evaluations ✓</td>
<td></td>
</tr>
<tr>
<td>Cross-lingual topic modeling</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>Datasets ✓ Models ✓ Evaluations ✓</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison between the latest OCTIS (Terragni et al., 2021) and TOPMOST. Our TOPMOST covers more topic modeling scenarios and their corresponding datasets, models, and evaluations.

1 Introduction

Topic models have been a fundamental and prevalent research area for decades. They aim to understand documents in an unsupervised fashion by discovering latent topics from them and inferring their topic distributions (Churchill and Singh, 2022b). Besides the basic topic modeling scenario (Blei et al., 2003), various other scenarios have been explored, e.g., hierarchical, dynamic, and cross-lingual topic modeling (Griffiths et al., 2003; Blei and Lafferty, 2006; Mimno et al., 2009). Current topic models can be categorized into two types. The first type is conventional topic models which follow either non-negative matrix factorization (Lee and Seung, 2000; Kim et al., 2015; Shi et al., 2018) or probabilistic graphical models via Markov Chain Monte Carlo (Steyvers and Griffiths, 2007) or Variational Inference (Blei et al., 2017). The second type is recently popular neural topic models, learned through gradient back-propagation (Zhao et al., 2021a; Wu et al., 2024b). Thus they can avoid the laborious model-specific derivations of conventional models, attracting more research attention. Due to the effectiveness and interpretability of topic models, they have inspired various downstream tasks and applications, e.g., text analysis and content recommendation (Boyd-Graber et al., 2017). Despite these significant achievements, quick utilization and fair comparisons of various topic models remain a formidable challenge. The reason lies in their unsystematic model implementations as well as inconsistent dataset and evaluation settings across papers, even within a paper (Hoyle et al., 2021).

Several topic modeling toolkits emerge in response to this challenge by integrating different
topic models and evaluations. However, they fail to fully meet practical requirements due to lacking certain essential features. Early toolkits (McCallum, 2002; Rehurek and Sojka, 2011; Qiang et al., 2020; Lisena et al., 2020) often lack the support for neural topic models or necessary steps in the topic modeling lifecycle, e.g., data preprocessing and evaluations. The latest toolkit OCTIS (Terragni et al., 2021) is more comprehensive, but as shown in Table 1 and Figure 1, it solely considers basic and hierarchical topic modeling scenarios and overlooks the latest advancements of neural topic models, offering only two neural topic models introduced after 2018. As a consequence, these issues pose hurdles to the comparisons, developments, and applications of topic models.

To resolve these issues, we in this paper introduce Topic Modeling System Toolkit (TopMost), which supports extensive features. In contrast to existing toolkits, TopMost thoroughly incorporates the most prevalent topic modeling scenarios: basic, hierarchical, dynamic, and cross-lingual topic modeling, as well as the latest neural topic models as detailed in Table 1 and Figure 1. It covers the entire lifecycles of these scenarios, including datasets, preprocessing, models, training, and evaluations. More importantly, TopMost adheres to an object-oriented paradigm with a highly cohesive and decoupled modular design. This enhances the readability and extensibility of TopMost, enabling users to flexibly customize their own datasets, models, and evaluations for their diverse research or application purposes. As a result, TopMost excels in fulfilling the practical requirements of topic modeling. We conclude the advantages of our TopMost as follows:

- TopMost provides handy and complete cutting-edge topic models for various scenarios;
- TopMost allows users to effortlessly and fairly compare topic models through comprehensive evaluation metrics;
- TopMost with better readability and extensibility facilitates the smooth development of new topic models and downstream applications.

2 Related Work

Throughout the long history of topic modeling, numerous toolkits have emerged and gained widespread adoption. The earliest among those include Mallet (McCallum, 2002)1 and gensim (Rehurek and Sojka, 2011)2. While these fundamental frameworks sufficiently embrace conventional topic models, they generally overlook the recent advancements in neural topic models. STTM (Qiang et al., 2018) particularly focuses on probabilistic short text topic models, like BTM (Yan et al., 2013) and DMM (Yin and Wang, 2014). A more recent entrant, OCTIS (Terragni et al., 2021), integrates both conventional and neural topic models. Nevertheless, it merely covers basic and hierarchical topic modeling scenarios and neglects the latest neural topic models developed after 2018. Moreover, OCTIS couples the implementations of model construction and training, exacerbating the challenges of toolkit maintenance. Different from these existing work, our TopMost extensively incorporates a spectrum of popular topic modeling scenarios and the latest developments in neural topic models. In addition, our TopMost clearly

---

1https://mimno.github.io/Mallet/topics.html
2https://radimrehurek.com/gensim/
Table 2: Summary of topic modeling scenarios, topic models, evaluation metrics, and datasets covered by TOPMOST.

decouples each step (data, models, and training) in the topic modeling lifecycles, resulting in neat code structures and simplified maintenance.

3 Overview of Toolkit Design and Architecture

In this section, we delineate the overview of our toolkit design and architecture. We build TOPMOST with Python and use PyTorch (Paszke et al., 2019) as the neural network framework for neural topic models. Figure 2 illustrates the overall architecture of TOPMOST.

3.1 Topic Modeling Scenarios and Topic Models

As summarized in Table 2, TOPMOST reaches a wider coverage by involving the 4 most popular topic modeling scenarios and their corresponding conventional or neural topic models.

Basic Topic Modeling discovers a number of latent topics from normal documents like news articles and web snippets, as the most common scenario (Blei et al., 2003). For basic topic models, TOPMOST supports conventional LDA (Blei et al., 2003), NMF (Lee and Seung, 2000), and most of the mainstream neural models such as ProdLDA (Srivastava and Sutton, 2017), ETM (Dieng et al., 2020), CombinedTM (Bianchi et al., 2021), BERTopic (Grootendorst, 2022), TSCTM (Wu et al., 2022), ECRTM (Wu et al., 2023b), and FASTopic (Wu et al., 2024c).

Hierarchical Topic Modeling organizes topics into a tree structure instead of flat topics in the basic topic modeling (Griffiths et al., 2003; Isonuma et al., 2020). Topics at each level of the structure involve different semantic granularity: child topics are more specific to their parent topics. This provides more desirable granularity for downstream applications. Hierarchical topic models in TOPMOST include conventional HDP (Teh et al., 2006) and recently popular neural hierarchical topic models, e.g., HyperMiner (Xu et al., 2022), ProGBN (Duan et al., 2023), and TraCo (Wu et al., 2024d).

Dynamic Topic Modeling discovers the evolution of topics in sequential documents, such as the conference papers published by year (Blei and Lafferty, 2006). This discloses how topics emerge, grow, and decline over time due to real-world trends and events, which has derived applications like trend analysis and public opinion mining (Li et al., 2020; Churchill and Singh, 2022a). For dynamic topic models, we provide the conventional
DTM (Blei and Lafferty, 2006) and its neural variant, DETM (Dieng et al., 2019). We also cover recent CFDTM (Wu et al., 2024a).

**Cross-lingual Topic Modeling** discovers aligned cross-lingual topics from bilingual corpora (Mimno et al., 2009). These reveal the commonalities and differences across languages and cultures, enabling cross-lingual text analysis without supervision (Yuan et al., 2018; Yang et al., 2019). Cross-lingual topic models in TOPMOST include NMTM (Wu et al., 2020a) and InfoCTM (Wu et al., 2023a).

We carefully adapt the original implementations of these topic models and unify their APIs of initialization, training, and testing, ensuring that our toolkit remains user-friendly, readable, and extendable. Note that we will constantly update TOPMOST to include more newly released models.

### 3.2 Datasets and Preprocessing

TOPMOST contains extensive benchmark datasets for the involved topic modeling scenarios, as reported in Table 2. We summarize the statistics of these datasets in Tables 3 to 5.

For basic and hierarchical topic modeling, we have the following datasets: (i) **20NG** (20 News Groups, Lang, 1995) is one of the most widely used datasets for evaluating topic models, including news articles with 20 labels. (ii) **IMDB** (Maas et al., 2011) is the movie reviews from the IMDB website, containing two sentimental labels, positive and negative. (iii) **Wikitext-103** (Merity et al., 2016) includes Wikipedia articles (Nguyen and Luu, 2021).

For dynamic topic modeling, TOPMOST provides the datasets as (i) **NeurIPS** includes the published papers at the NeurIPS conference from 1987 to 2017. (ii) **ACL** (Bird et al., 2008) is an article collection between 1973 and 2006 from ACL Anthology. (iii) **NYT** contains the news articles in the New York Times, from 2012 to 2022, with 12 categories, like “Arts”, “Business”, and “Health”.

For cross-lingual topic modeling, we offer the following bilingual datasets: (i) **ECNews** (Wu et al., 2020a) is a collection of English and Chinese news with 6 categories like business, education, and entertainment. (ii) **Amazon Review** (Wu et al., 2020a) includes English and Chinese reviews from the Amazon website, where each review has a rating from one to five. We simplify it as a binary classification task by labeling reviews with ratings of five as “1” and the rest as “0” following Yuan et al. (2018). (iii) **Rakuten Amazon** (Wu et al., 2023a) contains Japanese reviews from Rakuten (a Japanese online shopping website, Zhang and LeCun, 2017), and English reviews from Amazon (Yuan et al., 2018). Similarly, it is also simplified as a binary classification task according to the ratings. Note that basic topic models can employ the datasets for dynamic topic modeling as well.

We preprocess these datasets with standard steps, such as removing stop words and punctuation, removing short tokens, and filtering low-frequency words (Card et al., 2018; Wu et al., 2020b). Users can directly download these off-the-shelf datasets for experiments through TOPMOST from our GitHub repository. See Appendix A for more details of these datasets. We also provide configurable preprocessing implementations, allowing users to flexibly customize their datasets.

### 3.3 Evaluation Metrics

TOPMOST provides sufficient evaluation metrics to evaluate topic models. We first evaluate the quality of discovered topics in terms of topic coherence (TC, Newman et al., 2010) and topic diversity (TD, Dieng et al., 2020). TC refers to the coherence between the top words of discovered topics, and TD measures the differences between topics. We consider different implementations of TC and TD, for example, NPMI (Lau et al., 2014), $C_V$ (Röder et al., 2015), and TU (Nan et al., 2019), for extensive comparisons.

Then, we evaluate the quality of inferred doc-topic distributions via extrinsic tasks: text classification and clustering (Wu and Li, 2019; Zhao et al., 2021b; Nguyen et al., 2024). For classification, we train an ordinary classifier (e.g., SVM) with doc-topic distributions as document features and predict the labels of others. For clustering, we use the most significant topics in doc-topic distributions as clustering assignments.

Apart from these fundamental ones, we addition-

---

1[^1]: https://ai.stanford.edu/~amaas/data/sentiment/ aclImdb_v1.tar.gz
3[^3]: https://www.kaggle.com/datasets/benhamner/nips-papers
4[^4]: https://aclanthology.org/
5[^5]: https://huggingface.co/datasets/Matthewww/nyt_news
6[^6]: https://github.com/bobxwu/NMTM
ally include metrics for special scenarios. For cross-lingual topic modeling, we measure the average TD over all languages and evaluate the alignment between cross-lingual topics with cross-lingual NPMI (CNPMI, Hao and Paul, 2018). For hierarchical topic modeling, we evaluate the quality of discovered topic hierarchies, concerning the coherence and diversity between parent and child topics, the diversity between parent and non-child topics, and the diversity between sibling topics (Chen et al., 2021b,a; Wu et al., 2024d).

4 Comparison to Existing Toolkit

To highlight our significant strengths, we compare our TopMOST with the latest counterpart, OCTIS (Terragni et al., 2021), which integrates more features than earlier toolkits. Our TopMOST outperforms OCTIS in three key aspects:

(i) As detailed in Table 1, TopMOST offers a broader coverage of topic modeling scenarios, accompanied by corresponding datasets, models, and evaluation metrics. This better fulfills the various requirements of researchers and developers.

(ii) TopMOST provides a more extensive array of topic models compared to OCTIS. As reported in Figure 1 while OCTIS merely includes 4 neural topic models, TopMOST incorporates 16 ones, including the latest NSTM (Zhao et al., 2021b), HyperMiner (Xu et al., 2022), and ECRTM (Wu et al., 2023b). These advanced models empower users with cutting-edge topic modeling techniques and simplify their comparisons and applications.

(iii) TopMOST entirely decouples the implementations of data loading, model construction, model training, and evaluations, as illustrated in Figure 2. This design streamlines the code structure for high reusability and facilitates fair comparisons among diverse topic models. It aligns with prominent libraries such as Huggingface Transformers and PyTorch Lightning. See the code examples in Sec. 5.

5 Toolkit Usage

We showcase the simplicity and user-friendly design of our TopMOST toolkit with code examples. Users can directly install our TopMOST through pip:

```
$ pip install topmost
```

Figure 3 shows how to quickly utilize TopMOST to discovers topics from documents with a few handy steps: dataset preprocessing, model construction (here ProdLDA (Srivastava and Sutton, 2017)), and training. We emphasize that our TopMOST supports other languages besides English. We can simply employ different tokenizers in the preprocessing for other languages, for example, jieba\(^{10}\) for Chinese and nagisa\(^{11}\) for Japanese. Other preprocessing settings are also configurable, including maximum vocabulary size, stop words, and maximum or minimum document frequency. This allows users to flexibly apply our toolkit.

```
from topmost.data import RawDatasetHandler
from topmost.models import ProdLDA
from topmost.trainers import BasicTrainer

docs = [ "A document about space, satellite, launch, orbit.", # more example documents... ]
# build a dataset
dataset = RawDatasetHandler(docs)
# create a topic model
model = ProdLDA(dataset.vocab_size)
# create a trainer
trainer = BasicTrainer(model)
topic_top_words, doc_topic_dist = trainer.fit_transform(dataset)
```

Figure 3: A code example for quick start.

Figure 4 exemplifies how to train a topic model with preprocessed datasets. The training of other topic models follows similar steps.

```
from topmost.data import download_dataset, BasicDatasetHandler
from topmost.models import ProdLDA
from topmost.trainers import BasicTrainer

download_dataset(‘20NG’, cache_path=’./datasets’)  # load a dataset
dataset = BasicDatasetHandler(’./datasets/20NG’)  # create a topic model
model = ProdLDA(dataset.vocab_size)
# create a trainer
trainer = BasicTrainer(model)
trainer.train(dataset)
```

Figure 4: A code example for training a topic model (ProdLDA (Srivastava and Sutton, 2017)).

Figure 5 shows how to fully evaluate the trained topic model with diverse metrics including topic coherence, topic diversity, text classification, and text clustering.

\(^{10}\)https://github.com/fxsjy/jieba

\(^{11}\)https://github.com/taishi-i/nagisa
Figure 6: Demonstration of testing new documents. It plots the inferred topic distribution of an input document from a trained topic model.

```
from topmost.evaluations import compute_topic_diversity,
compute_topic_coherence, evaluate_clustering,
evaluate_classification

# doc-topic distributions
train_theta, test_theta = trainer.export_theta(dataset)

# top words of topics
topic_top_words = trainer.export_top_words(dataset.vocab)

# topic coherence
calculate_topic_coherence(topic_top_words, dataset.train_texts)

# topic diversity
calculate_topic_diversity(topic_top_words)

# text clustering
evaluate_clustering(test_theta, dataset.test_labels)

# text classification
evaluate_classification(train_theta, test_theta, dataset.train_labels, dataset.test_labels)
```

Figure 5: A code example for evaluating a topic model, including topic coherence, topic diversity, text classification, and clustering.

The above examples illustrate that TopMost clearly decouples the APIs of data, models, and training, so TopMost becomes more accessible, easy-to-use, and extendable to users. Due to limited page space, see more examples and tutorials on our GitHub project page, like data preprocessing and other topic modeling scenarios.

Figure 7: Visualization of discovered topics. It plots the top related words of each topic and the modeled word distributions.

6 Visualization Interfaces

TopMost furthermore provides visualization interfaces for topic models. We create a web demo system with Flask 12 as the server framework following Material design to visualize and test topic models. It is designed to be intuitive and user-friendly, enabling users to easily understand and leverage topic models.

Figure 7 shows the visualization of topics. We plot the top related words of discovered topics and the modeled probability of each word. For example,

---

12https://flask.palletsprojects.com/
Topic#15 in Figure 7 mostly relates to words like “matlab”, “plot”, and “vectors”. By selecting the index or clicking the Previous and Next buttons, we can view the details of any topic.

Figure 6 demonstrates the interactive utilization of a trained topic model. Upon inputting a document, we can click the upload button to obtain the inferred topic distribution of the document. The horizontal bar chart in Figure 6 plots the distribution over all topics of the input How to plot lines between all points in vector in Matlab?. We see that the topic distribution mainly lies on Topic#15, which refers to Matlab.

7 Conclusion and Future Work

In this paper, we present TopMOST, an open-source, comprehensive, and up-to-date topic modeling system toolkit. TopMOST provides complete lifecycles of various topic modeling scenarios, including datasets, preprocessing, models, training, and evaluations, which outperforms existing counterparts. TopMOST allows users to smoothly explore topic models, verify their new ideas, and develop novel topic modeling applications. This benefits both the communities in academia and industry. In the future, we plan to keep TopMOST updated to incorporate more latest topic models and support more features to facilitate the research and application of topic modeling.

Limitations

We consider the following limitations of TopMOST. First, TopMOST only includes the mainstream evaluation metrics. Some less popular ones like perplexity are ignored. Second, TopMOST does not cover the topic models based on prompting large language models (Pan et al., 2023; Wu et al., 2024e; Pham et al., 2023). Different from LDA-like models, they define a topic as a textual description, so we cannot assess them through existing evaluation metrics.

References


Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pages 142–150.


Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Topic modeling as multi-objective optimization with setwise contrastive learning. In The Twelfth International Conference on Learning Representations.


high-performance deep learning library. Advances in neural information processing systems, 32.


Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xingdong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering, 34(3):1427–1445.


Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024b. A survey on neural topic models: Methods, applications, and challenges. Artificial Intelligence Review.


### Table 3: Statistics of pre-processed datasets for cross-lingual topic modeling.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language</th>
<th>#docs</th>
<th>Vocabulary Size</th>
<th>Average length</th>
<th>#labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECNews English</td>
<td>English</td>
<td>46,870</td>
<td>5,000</td>
<td>12.0</td>
<td>6</td>
</tr>
<tr>
<td>ECNews Chinese</td>
<td>Chinese</td>
<td>50,000</td>
<td>5,000</td>
<td>10.6</td>
<td></td>
</tr>
<tr>
<td>Amazon Review</td>
<td>English</td>
<td>25,000</td>
<td>5,000</td>
<td>30.6</td>
<td>2</td>
</tr>
<tr>
<td>Amazon Review Chinese</td>
<td></td>
<td>25,000</td>
<td>5,000</td>
<td>43.2</td>
<td></td>
</tr>
<tr>
<td>Rakuten Amazon</td>
<td>English</td>
<td>25,000</td>
<td>5,000</td>
<td>30.6</td>
<td>2</td>
</tr>
<tr>
<td>Rakuten Amazon Japanese</td>
<td></td>
<td>25,000</td>
<td>5,000</td>
<td>22.5</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Statistics of pre-processed datasets for basic and hierarchical topic modeling.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#docs</th>
<th>Vocabulary Size</th>
<th>Average Length</th>
<th>#labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>18,846</td>
<td>5,000</td>
<td>110.5</td>
<td>20</td>
</tr>
<tr>
<td>IMDB</td>
<td>50,000</td>
<td>5,000</td>
<td>95.0</td>
<td>2</td>
</tr>
<tr>
<td>Wikitext-103</td>
<td>28,532</td>
<td>10,000</td>
<td>1,355.4</td>
<td>/</td>
</tr>
</tbody>
</table>

### Table 5: Statistics of pre-processed datasets for dynamic topic modeling.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#docs</th>
<th>Vocabulary Size</th>
<th>Average Length</th>
<th>#labels</th>
<th>#time slices</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeurIPS</td>
<td>7,237</td>
<td>10,000</td>
<td>2,085.9</td>
<td>/</td>
<td>31</td>
</tr>
<tr>
<td>ACL</td>
<td>10,560</td>
<td>10,000</td>
<td>2,023.0</td>
<td>/</td>
<td>31</td>
</tr>
<tr>
<td>NYT</td>
<td>9,172</td>
<td>10,000</td>
<td>175.4</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

### A Datasets

Tables 3 to 5 report the statistics of datasets for different topic modeling scenarios after preprocessing. Users can directly download all these datasets via TOPMOST from our GitHub repository.