STREAM-ZH: Simplified Topic Retrieval Exploration and Analysis Module for Chinese Language

Hongyi Li[1,2*], **Jianjun Lian**[2,3*], **Anton Frederik Thielmann**[4†], **Andre Python**[2,5,6‡]
[1]School of Mathematical Sciences, Zhejiang University
[2]Center for Data Science, Zhejiang University
[3]International Business School, Zhejiang University; [4]Amazon Music
[5]School of Medicine, Zhejiang University
[6]Centre for Human Genetics, Nuffield Department of Medicine, Oxford University

Abstract

We introduce Simplified Topic Retrieval Exploration and Analysis Module for Chinese language (STREAM-ZH), the first topic modeling package to fully support the Chinese language across a broad range of topic models, evaluation metrics, and preprocessing workflows. Tailored to both simplified and traditional Chinese language, our package extends the STREAM topic modeling framework with a curated collection of preprocessed textual datasets in Chinese from which we assess the performance of classical, neural, and clustering topic models using commonly-used intruder, diversity, and coherence metrics. The results of a benchmark analysis bring evidence that within our framework, topic models may generate coherent and diverse topics from datasets in Chinese language, outperforming those generated by topic models using English-translated textual input. Our framework facilitates multilingual accessibility and research in topic modeling applied to Chinese textual data. The code is available at the following link: https://github.com/AnFreTh/STREAM

1 Introduction

Topic modeling is a key tool in the field of automatic text analysis, which aims to automatically identify potential topics and infer their distribution from large-scale unstructured textual data through unsupervised learning methods (Blei et al., 2003; Abdelrazek et al., 2023; Churchill and Singh, 2022). Topic modeling has been widely used in different fields (Liu et al., 2016; Hong and Davison, 2010; Girdhar et al., 2013) and has provided effective support for natural language processing (NLP) tasks such as text classification and sentiment analysis (Boyd-Graber et al., 2017; Abdelrazek et al., 2023).

Recent studies (Pham et al., 2023; Wang et al., 2023; Meram et al., 2025) have explored the use of large language models (LLMs) for topic modeling, leveraging their strong contextual understanding to generate more coherent and interpretable topics. However, LLM-based approaches face several practical challenges. They are not plug-and-play, often requiring multiple iterative steps and carefully crafted prompts (Pham et al., 2023; Mu et al., 2024); they are constrained by limited context windows (Pham et al., 2023), which prevents them from capturing global topic structures; and they incur substantial computational overhead (Wang et al., 2023). Moreover, open-source LLMs often struggle to follow instructions reliably (Isonuma et al., 2024), while closed-source LLMs introduce significant costs and potential privacy risks (Pham et al., 2023), limiting their scalability in real-world applications.

Several open-source toolkits have been released to meet practical needs. For example, Topic Modeling System Toolkit (TopMost) (Wu et al., 2023) and Simplified Topic Retrieval Exploration and Analysis Module (STREAM) (Thielmann et al., 2024b), cover the entire life cycle of topic modeling, including data preprocessing, model training, and topic evaluation.

While current toolkits satisfy the needs of users in a wide range of research and applications, they are built for English and other Latin-alphabet languages (McCallum, 2002; Řehřek et al., 2011; Qiang et al., 2018; Terragni et al., 2021a). One may translate text in non-Latin language (such as Japanese, Korean, and Chinese) into English before topic modeling. However, this two-step process may be time consuming, and more problematically, lead to incoherent theme semantics (Zhao et al., 2011). This is partially due to the fact that these languages lack of clear word separators, making a direct integration into a Latin alphabet-based toolkits challenging. Our experiment (Table 2) brings

*These authors contributed equally to this work.
†Work done while at Technical University Clausthal.
‡Corresponding author: apython@zju.edu.cn

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 371–383
March 24-29, 2026 ©2026 Association for Computational Linguistics

further evidence of suboptimal results in a case study using Chinese text translated into English as input for topic modeling.

Familia (Jiang et al., 2021), an industrial-grade topic modeling framework, supports topic inference and semantic matching computation for Chinese corpora. However, despite its high computational efficiency, the open-source version of Familia [1] considers only three topic models: Latent Dirichlet Allocation (LDA) (Blei et al., 2003), SentenceLDA, and Topical Word Embedding (TWE), and lacks data preprocessing and topic evaluation modules. To satisfy the expected needs of a large number of current and potential future users—around 1.5 billion people speak Chinese across the world—the STREAM-ZH software package fills an important gap by offering a toolkit specifically designed for the Chinese language that covers a wide range of possible applications.

1.1 Contributions

Our contributions can be summarized as follows:

- We present the first topic modeling package with comprehensive support for Chinese, covering a wide range of models, evaluation metrics, and preprocessing pipelines.
- The toolkit supports flexible Chinese preprocessing, including word segmentation (with multiple tools), custom dictionaries, and POS tagging.
- Four curated and preprocessed Chinese textual datasets are included as benchmarks for topic modeling.
- We benchmark several standard topic models on these datasets using intruder detection, diversity, and coherence metrics.

2 Simplified Topic Retrieval Exploration and Analysis Module for Chinese

To the best of our knowledge, STREAM-ZH is the first topic modeling toolkit that fully supports Chinese in dataset preprocessing, topic modeling of a broad range of models and evaluation metrics. It fills the gap in the processing capabilities of existing topic modeling toolkits and marks an important step forward in the multilingual applicability of topic modeling.

[1] https://github.com/baidu/Familia?tab=readme-ov-file

2.1 Available Datasets

To complement the English-only datasets included in STREAM, we added four Chinese datasets:
THUCNews[2] (Sun et al., 2016), generated by filtering Sina News RSS feeds (2005–2011), and **THUCNews_small**, a subset with 1,000 documents per category;
FUDANCNews[3], originally for text classification, merged from its training and test sets;
TOUTIAO[4], a headline dataset collected in May 2018, and **TOUTIAO_small**, containing 1,400 documents per category;
CMtMedQA_ten, derived from CMtMedQA[5] (Yang et al., 2024), a Chinese multi-round medical conversation corpus, by selecting ten medical themes, with a corresponding **CMtMedQA_small** subset of 1,000 documents per category.

Detailed statistics (e.g., number of documents, average document length) are provided in Table S1 in the appendix.

2.2 Preprocessing Domain-Specific text in Chinese language

Unlike English, Chinese text lacks natural word boundaries (Qin et al., 2016). Therefore, how to accurately segment a continuous sequence of Chinese characters into individual words has become a key issue in Chinese natural language processing. Moreover, in practical applications, texts in different domains usually have specific lexical rules, and the use of generalized Chinese Word Segmentation (CWS) tools may lead to a decrease in the lexical accuracy of domain-specific texts (Luo et al., 2019), which in turn affects the effectiveness of topic modeling. To better accommodate the needs of users for domain-specific text segmentation, STREAM-ZH provides five mainstream open-source CWS tools with excellent performance, including Jieba[6], PKUSeg[7] (Luo et al., 2019), THULAC, HanLP[8] (He and Choi, 2021), and SnowNLP[9], adapted to various data characteristics. We do not consider CWS methods based on

[2] https://github.com/thunlp/THULAC-Python
[3] https://gitcode.com/open-source-toolkit/6a679
[4] https://github.com/aceimnorstuvwxz/TOUTIAO-text-classfication-dataset
[5] https://huggingface.co/datasets/Suprit/CMtMedQA
[6] https://github.com/fxsjy/Jieba
[7] https://github.com/lancopku/PKUSeg-python
[8] https://github.com/hankcs/HanLP
[9] https://github.com/isnowfy/SnowNLP

pre-trained language models such as CWSeg (Li et al., 2023), as they are typically not computationally efficient for the large-scale textual datasets common in topic modeling.

3 Results

In the following we report the results of experiments using STREAM-ZH to compare topic models based on the newly introduced Chinese datasets (THUCNews_small, FUDANCNews, TOUTIAO_small and CMtMedQA_small). The first three datasets belong to the news domain. The FUDANCNews dataset has the longest average document length, the THUCNews_small dataset has a moderate document length, and the TOUTIAO_small dataset consists of news headlines, which are considerably shorter. The CMtMedQA_small dataset is a medical domain dataset obtained from real doctor-patient exchanges and contains a large number of active inquiry statements.

We consider topic modeling evaluation metrics in STREAM, including intruder metrics (Thielmann et al., 2024a)(Average Intruder Similarity (ISIM), Intruder Shift (ISH), Intruder Accuracy (INT)), diversity metrics (Word Embedding-based Weighted Sum Similarity (WESS) and Topic Expressivity (EXPRS)), Coherence Metrics (Embedding Coherence (COH) (Terragni et al., 2021b) and Normalized pointwise mutual information (NPMI) (Lau et al., 2014)). We consider three classes of topic models such as classical, neural, and clustering topic models. Specifically, we trained LDA (Gensim[10]), Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000), Contextualized Topic Models (CTM) (Bianchi et al., 2020), BERTopic (Grootendorst, 2022) and KmeansTM (Thielmann et al., 2024a).

3.1 Experimental evaluation of topic models on news datasets

Table 1 shows that on the THUCNews_small dataset, KmeansTM performs best on the intruder and diversity metrics, and scores high on the coherence metrics, indicating that its topics are relatively well-separated and less redundant. NMF scores significantly higher on the coherence metrics than the other models, which suggests that it generates semantically coherent topics. These results highlight a clear trade-off: NMF prioritizes coherence, while KmeansTM better preserves topic separation and diversity.

NMF and KmeansTM achieve the best performance with regard to intruder, diversity, and coherence metrics on the FUDANCNews dataset, indicating they generate both coherent and diverse topics. On TOUTIAO_small (short texts) dataset, neural topic modeling becomes more competitive: CTM achieves best on nearly all three metrics. This aligns with the intuition that contextual embeddings can better exploit limited word co-occurrence in sparse documents, improving both topic discriminability and semantic consistency. Notably, NMF attains the highest NPMI but lags behind CTM and KmeansTM on diversity, suggesting that high coherence does not necessarily imply non-redundant topics under short-text conditions.

3.2 Experimental evaluation of topic models on a medical dataset

The results (Table 1) on CMtMedQA_small dataset outside journalistic domain show that KmeansTM performs best in all evaluation metrics, especially with regard to the metric "INT", suggesting that the topics it generates are particularly outstanding in terms of distinctiveness, in addition to being coherent and diverse. A consistent pattern across datasets is that LDA underperforms on all metric families, suggesting that in Chinese setting it struggles to produce both semantically coherent and discriminable topics. Overall, KmeansTM systematically performs better on all four datasets used in this study, indicating that the clustering-based approach transfers robustly beyond journalistic content, producing topics that are simultaneously coherent and separable.

Qualitative examples of topic modeling results on the four datasets are provided in the appendix Tables S3, S4, S5, and S6. Additionally, we hypothesize that the choice of CWS tools may significantly impact topic modeling performance in domain-specific datasets. To verify this, we conducted a Friedman test on the CMtMedQA_small dataset results. The test revealed statistically significant differences across CWS tools in the intruder and coherence metrics, whereas no significant variation was observed for diversity metrics. Consequently, we performed a post-hoc Nemenyi test to examine specific pairwise differences, confirming that the choice of CWS tool significantly influences the quality of the generated topics. Detailed experimental results and statistical tests can be found

[10] https://github.com/piskvorky/gensim

Metrics	Intruder			Diversity		Coherence	
Model	ISIM ↓	INT ↑	ISH ↓	WESS ↓	EXPRS ↓	NPMI ↑	COH ↑
Dataset: THUCNews_small							
LDA	0.577	0.143	0.719	0.897	0.922	-0.649	0.597
NMF	0.517	**0.505**	0.625	0.745	0.838	**0.508**	**0.647**
CTM	0.499	0.498	0.609	0.732	0.829	0.442	0.633
BERTopic	0.509	0.392	0.620	0.747	0.833	0.426	0.630
KmeansTM	**0.486**	0.494	**0.601**	**0.717**	**0.824**	0.402	0.628
Dataset: FUDANCNews							
LDA	0.524	0.231	0.672	0.843	0.870	-0.216	0.565
NMF	**0.502**	0.366	0.628	0.762	0.839	**0.362**	0.608
CTM	0.517	0.437	0.631	0.763	0.842	0.326	0.639
BERTopic	0.528	0.239	0.655	0.798	0.855	0.219	0.602
KmeansTM	0.512	**0.450**	**0.626**	**0.751**	**0.835**	0.314	**0.643**
Dataset: TOUTIAO_small							
LDA	0.546	0.131	0.705	0.881	0.914	-0.702	0.562
NMF	0.527	0.270	0.667	0.819	0.859	**0.045**	0.598
CTM	**0.471**	0.567	**0.565**	**0.663**	**0.782**	-0.112	**0.667**
BERTopic	0.608	0.175	0.744	0.892	0.912	-0.254	0.632
KmeansTM	0.493	**0.585**	0.592	0.709	0.816	-0.062	0.664
Dataset: CMtMedQA_small							
LDA	0.645	0.105	0.784	0.941	0.912	-0.128	0.637
NMF	0.616	0.211	0.735	0.862	0.874	0.241	0.674
CTM	**0.620**	0.242	0.732	0.854	0.871	0.225	0.686
BERTopic	0.632	0.392	0.733	0.827	0.842	0.136	0.724
KmeansTM	**0.620**	**0.592**	**0.710**	**0.788**	**0.811**	**0.254**	**0.743**

Table 1: Benchmark results on the introduced Chinese datasets. All models and evaluation metrics use the Conan-embedding-v1 pre-trained embedding model (Li et al., 2024) when applicable. Model hyperparameters are tuned over 20 trials. The intruder metrics are averaged over 100 runs. All models, except BERTopic, are fitted with a pre-specified number of 14, 20, 14, and 10 topics, respectively. BERTopic automatically detects the optimal number of topics, hence we extract the corresponding number of topics from the output in order. The ↑ symbol indicates that a higher value is better, while the ↓ symbol indicates that a lower value is better.

in the appendix Tables S7 and S8. Note that the Nemenyi test was employed instead of pairwise t-tests to strictly control for the family-wise error rate in this multi-method comparison context.

3.3 Sensitivity analyses

We further assess robustness to two practical perturbations: class imbalance and the number of topics. First, we construct three imbalanced variants of the datasets via stratified sampling: THUCNews_imbalance, TOUTIAO_imbalance, and CMtMedQA_imbalance (category-wise document counts are reported in the Table S9). By rerunning the same benchmarking pipeline, we show that the results are mainly stable (see Table S10): KmeansTM remains competitive and typically leading on intruder and diversity metrics across imbalanced datasets, while CTM remains strong on short-text dataset (TOUTIAO_imbalance) and is highly competitive on CMtMedQA_imbalance. The leading method on coherence can shift under imbalance (e.g., on THUCNews_imbalance, BERTopic improves and becomes best on coherence metrics), suggesting that coherence metrics are more sensitive to distributional skew than intruder and diversity metrics under our evaluation setup. Compared to the results of the original dataset (Table 1), we see modest decrease in overall topic quality for various models and datasets, confirming the fact that topic models usually obtain higher performance for balanced datasets.

Second, we evaluate sensitivity to the topic count by setting a range of values for the hyperparameter $K \in \{10, 30, 50, 100\}$ for KmeansTM on THUCNews_small (see Table S11). As K increases, intruder and diversity metrics degrade monotonically, while COH improves modestly. This behavior is consistent with over-segmentation: larger K can yield narrower topics that appear internally coherent, yet become increasingly fragmented and redundant as similar semantics are split across multiple nearby topics, thereby harming distinctiveness and diversity.

Overall, these sensitivity experiments indicate that the relative strengths observed in the main benchmark are not artifacts of a single sampling regime or a single choice of K, while also clarifying the coherence and diversity trade-off induced by increasing topic granularity.

3.4 Experimental evaluation of topic models relative to a machine translation approach

One may reasonably wonder whether a native incorporation of a non-Latin language for topic modeling, as proposed in our software STREAM-ZH, is justified given the high performance of current translation tools. In principle, text from any language can be translated into English and used in English-based topic modeling frameworks. To address this question, we compared the performance of STREAM-ZH based on original Chinese text with STREAM (Thielmann et al., 2024b) using English translations via Google translate. The Chinese topics obtained in the original Chinese datasets (using STREAM-ZH) will be compared with the Chinese translations of the English topics obtained in the English translation of the same datasets (using STREAM). The quality of these topics is evaluated using the previously introduced metrics. Metrics that depend on reference corpus are excluded since it would be unfair to use the original dataset as baseline. The results (Table 2) indicate that overall the topics generated directly using STREAM-ZH

Metrics	Intruder			Diversity		Coherence
	ISIM ↓	INT ↑	ISH ↓	WESS ↓	EXPRS ↓	COH ↑
THUCNews_small						
STREAM-ZH	**0.477****	**0.562*****	**0.582*****	**0.703*****	**0.812*****	**0.637*****
STREAM	0.487	0.460	0.601	0.729	0.824	0.617
FUDANCNews						
STREAM-ZH	**0.521***	0.442	0.638	**0.763****	**0.840*****	**0.642***
STREAM	0.528	0.445	0.642	0.772	0.848	0.638
TOUTIAO_small						
STREAM-ZH	**0.511***	**0.427*****	**0.619*****	**0.741*****	**0.825*****	**0.646*****
STREAM	0.518	0.359	0.638	0.777	0.848	0.617
CMtMedQA_small						
STREAM-ZH	0.615	**0.562*****	0.705	**0.796*****	**0.815*****	**0.741*****
STREAM	**0.604*****	0.380	0.704	0.808	0.825	0.711

Table 2: Benchmark results on the introduced Chinese datasets. All models use Conan-embedding-v1 for Chinese and paraphrase-MiniLM-L3-v2 (Reimers and Gurevych, 2019) for English, where applicable. We apply KmeansTM with fixed topic counts (14, 20, 14, or 10, depending on the dataset) to both original and translated texts, repeating each run 10 times. Metrics are averaged over runs. One-sided independent t-tests are used to determine whether observed differences are statistically significant. The best-performing model (higher or lower depending on the metric) is bolded. Significance levels: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

are more coherent and diverse. This finding suggests that topic modeling with native non-Latin language integration may be more effective than relying on translation-based approaches. This may be attributed to the fact that current processes of machine translation ineluctably reduce the information content of the original text such as ignoring semantic subtleties and cultural context (Naveen and Trojovskỳ, 2024; Yang et al., 2023), leading to suboptimal downstream performance. While part of this problem might be alleviated with the emergence of more powerful translation tools, intrinsic differences between languages still obstruct the complete transition of information to some extent. Although De Vries et al. (De Vries et al., 2018) reported that machine translation has only limited effects on bag-of-words-based topic modeling, it is important to note that their study focused on languages (Danish, German, Spanish, French, Polish) that are linguistically closer to English, and that they employed LDA, which showed the weakest performance in our benchmark experiments (Table 1).

4 Conclusion

In this paper, we propose STREAM-ZH, which extends the STREAM topic modeling framework to textual data in simplified and traditional Chinese. To address the challenges associated with the peculiarities of the Chinese language, we adapted the data processing flow so that users can independently choose suitable CWS tools, customize dictionaries, remove specific parts of speech as well as the conversion of text between traditional and simplified Chinese. In addition, we released four pre-processed Chinese datasets that can be used as benchmark datasets for Chinese topic modeling. The results of our experiments show that our software may assess the performance of topic models in various scenarios based on textual data in Chinese language. Additional experiments bring evidence that our approach outperforms English-based topic models using automatically translated textual data from Chinese to English. Future work may include exploring automated segmentation optimization, automatic selection of the most suitable CWS tool based on dataset characteristics, and extending the benchmark datasets to include more domain-specific corpora.

5 Limitations

Although STREAM-ZH is the first topic modeling framework that fully supports Chinese language, it still has several limitations and potential risks. First, our evaluation is based only on a limited number of Chinese benchmark datasets, which may not be sufficient to fully reflect the complexity of domain-specific (such as law, chemistry, etc.) corpora. Second, the framework currently does not support processing mixed Chinese and English texts, and can only perform topic modeling on data in a single language, which may limit its practicality in multilingual or cross-lingual scenarios. In addition, in the comparative experiment of topic modeling between original Chinese text and its English translated version, we only used Google Translate as the benchmark. Considering the rapid development of neural machine translation (NMT) tools, the gap between these two procedures may gradually narrow, and better results may be achieved using newer NMT tools. Potential risks include biases between automated evaluation metrics and human assessments that may lead to misinterpretation of the results, and ethical risks resulting from the openness of users to interpret the results of topic modeling.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFC2701905) and the National Natural

Science Foundation of China (T2350610281, 82273731, and 12531013). We thank the anonymous reviewers for their valuable suggestions.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Suman Adhya, Avishek Lahiri, Debarshi Kumar Sanyal, and Partha Pratim Das. 2022. Improving contextualized topic models with negative sampling. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 128–138.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

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

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

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

Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher. 2018. No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, 26(4):417–430.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

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

Yogesh Girdhar, Philippe Giguere, and Gregory Dudek. 2013. Autonomous adaptive underwater exploration using online topic modeling. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, pages 789–802. Springer.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Han He and Jinho D Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. *arXiv preprint arXiv:2109.06939*.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

Masaru Isonuma, Hitomi Yanaka, et al. 2024. Comprehensive Evaluation of Large Language Models for Topic Modeling. *arXiv preprint arXiv:2406.00697*.

Di Jiang, Yuanfeng Song, Rongzhong Lian, Siqi Bao, Jinhua Peng, Huang He, Hua Wu, Chen Zhang, and Lei Chen. 2021. Familia: A configurable topic modeling framework for industrial text engineering. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26*, pages 516–528. Springer.

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.

Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Qing Lei, Haifeng Li, and Yanxi Chen. 2021. How does Chinese segmentation strategy effect on sentiment analysis of short text? In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 234–240. IEEE.

Dedong Li, Rui Zhao, and Fei Tan. 2023. Cwseg: An efficient and general approach to Chinese word segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 1–10.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024. Conan-embedding: General text embedding with more and better negative samples. *arXiv preprint arXiv:2408.15710*.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5:1–22.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. PKUSEG: A toolkit for multi-domain Chinese word segmentation. *arXiv preprint arXiv:1906.11455*.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. *http://mallet. cs. umass. edu*.

Muhammet Bora Meram, Çağatay Kalkan, Tuğba Çelikten, and Aytuğ Onan. 2025. GPT vs. Other Large Language Models for Topic Modeling: A Comprehensive Comparison. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(3):116–130.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.

Palanichamy Naveen and Pavel Trojovskỳ. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Jipeng Qiang, Yun Li, Yunhao Yuan, Wei Liu, and Xindong Wu. 2018. STTM: A tool for short text topic modeling. *arXiv preprint arXiv:1808.02215*.

Zengchang Qin, Yonghui Cong, and Tao Wan. 2016. Topic modeling of Chinese language beyond a bag-of-words. *Computer Speech & Language*, 40:60–78.

Radim Řehřek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

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

Arik Reuter, Anton Thielmann, Christoph Weisser, Benjamin Säfken, and Thomas Kneib. 2025. Probabilistic topic modeling with transformer representations. *IEEE Transactions on Neural Networks and Learning Systems*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An efficient lexical analyzer for Chinese. *Retrieved Jan*, 10:2022.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *International conference on applications of Natural Language to information systems*, pages 33–45. Springer.

Anton Thielmann, Arik Reuter, Quentin Seifert, Elisabeth Bergherr, and Benjamin Säfken. 2024a. Topics in the haystack: Enhancing topic quality through corpus expansion. *Computational Linguistics*, 50(2):619–655.

Anton Thielmann, Arik Reuter, Christoph Weisser, Gillian Kant, Manish Kumar, and Benjamin Säfken. 2024b. STREAM: Simplified Topic Retrieval, Exploration, and Analysis Module. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 435–444.

Anton Thielmann, Christoph Weisser, Thomas Kneib, and Benjamin Säfken. 2023. Coherence based document clustering. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 9–16. IEEE.

Anton F Thielmann, Christoph Weisser, and Benjamin Säfken. 2024c. Human in the loop: How to effectively create coherent topics by manually labeling only a few documents per class. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8395–8405.

Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241. IEEE.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2023. Towards the TopMost: A topic modeling system toolkit. *arXiv preprint arXiv:2309.06908*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Yanxia Yang, Runze Liu, Xingmin Qian, and Jiayue Ni. 2023. Performance and perception: machine translation post-editing in Chinese-English news translation by novice translators. *Humanities and Social Sciences Communications*, 10(1):1–8.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2020. Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537*.

Qi Zhao, Zengchang Qin, and Tao Wan. 2011. What is the basic semantic unit of Chinese language? A computational approach based on topic models. In *Conference on Mathematics of Language*, pages 143–157. Springer.

Xinyue Zhao, Jianing Huang, Jing Zhang, and Yunsheng Song. 2024. The comprehensive analysis of the effect of Chinese word segmentation on fuzzy-based classification algorithms for agricultural questions. *International Journal of Fuzzy Systems*, 26(8):2726–2749.

A Chinese dataset preprocessing

Datasets Creating a custom dataset that includes the use of a specified Chinese Word Segmentation (CWS) tool and custom dictionary, the removal of a specified lexical property, and the conversion of traditional Chinese to simplified Chinese is embedded into one single function in python:

```
from stream.data_utils import TMDataset
df = pd.read_csv("your_data.txt")

dataset = TMDataset()
dataset.create_load_save_dataset(
    data=df,
    dataset_name="your_name",
    save_dir="save directory",
    doc_column="text", # column name
    where documents are stored
    label_column="labels",
    language = "chinese",
    stopwords_path = "your stopwords
    file path",
    segmentation_tool = "pkuseg", #
    Choose from jieba, pkuseg, hanlp,
    thulac, snownlp
    domain = 'medicine', # Only for "
    pkuseg", choose from news, web,
    medicine, tourism and mixed
    segmentation_dict = "userdict path",
    remove_pos = ['c', 'd'] # Remove
    conjunction and adverb
)
```

Preprocessing newly introduced Chinese datasets To mitigate potential errors and bias from the original Chinese language datasets, we carried out several preprocessing operations. From the raw data, we removed HTML tags, non-standard characters, numbers, punctuation marks, and English words to eliminate noise and retain core semantic information. We employed 'Jieba' for the segmentation process and filtered the segmentation results based on a predefined stopwords list to eliminate meaningless words. We further filtered out words composed of more than 10 characters. To maintain stability and representativeness of the lexical distribution, we set a minimum word frequency thresholds for each investigated dataset as follows: 200 for THUCNews, 100 for FUDANCNews, 20 for TOUTIAO and THUCNews_small, 10 for CMtMedQA_ten, 5 for TOUTIAO_small and CMtMedQA_small. The final preprocessing stage consisted of removing documents with empty lexical results to ensure the completeness and usability of the resulting pre-processed dataset. The statistical information of each pre-processed dataset can be found in Table S1.

B Evaluation Metrics and Hyperparameter Configuration

STREAM-ZH offers the following topic evaluation metrics. Intruder metrics evaluate the robustness of top words against intruder words (i.e., words that do not semantically belong to a given topic). Diversity Metrics evaluate how well topics are separated and how expressive they are. Coherence Metrics evaluate the internal semantic consistency of topics.

Intruder Metrics:

- ISIM: Measures the average cosine similarity between the embeddings of a topic's top words and the intruder word.
- INT: For a given topic and intruder word, this metric computes the fraction of cases where the intruder word is the least similar (in embedding space) to the top words of the topic.
- ISH: Quantifies the change in the centroid of a topic when an intruder word is inserted.

Diversity Metrics:

- Expressivity: Computes the cosine distance between a topic centroid and the centroid of meaningless words (e.g., stopwords).
- Word Embedding-based Weighted Sum Similarity: Evaluates the diversity of topics in the embedding space.

Coherence Metrics:

- Embedding Coherence: Calculates the cosine similarity between the centroid of topic words and the centroid of stopwords.
- NPMI: A classical coherence measure based on word co-occurrence statistics from the source corpus.

We also provide the hyperparameter configurations of all models (Table S2) in the benchmark experiments to better ensure reproducibility.

C Available Models

In addition to the models used in the experiments STREAM-ZH supports chinese topic models for the following classes: ETM (Dieng et al., 2020), DCTE (Thielmann et al., 2024c), CTMNeg (Adhya et al., 2022), ProdLDA and NeuralLDA (Srivastava and Sutton, 2017), TNTM (Reuter et al., 2025), NSTM (Zhao et al., 2020) and CBC (Thielmann et al., 2023).

Dataset	n	l	w	$tokens$
THUCNews	804,656	230.5 (168.7)	395,432	185,150,667
THUCNews_small	13,994	198.1 (158.6)	40,865	2,741,975
FUDANCNews	9,526	422.5 (340.3)	22,985	4,024,481
TOUTIAO	337,902	10.2 (4.8)	57,616	3,451,639
TOUTIAO_small	19,399	8.1 (4.1)	12,777	156,744
CMtMedQA_ten	48,413	166.1 (47.1)	22,404	8,042,781
CMtMedQA_small	9,909	164.6 (47.0)	12,885	1,631,438

Table S1: Statistics of the pre-processed Chinese datasets. The number of documents (n) refers to the total number of document samples in each dataset. The average document length (l) indicates the average number of words per document (with standard deviation in parentheses). The number of unique words (w) counts the number of distinct words appearing in the pre-processed dataset, and the number of tokens ($tokens$) represents the total token count in each datasets.

D Qualitative Example of Topic Outputs

To provide a more comprehensive view of the benchmark results, beyond the automatic evaluation metrics, we further present example topic outputs from multiple models on different corpora (Table S3, S4, S5, S6). These examples (with English translations) illustrate what kind of content the models are able to capture in the Chinese context.

E Evaluating the effect of CWS tools on topic modeling

Prior studies have shown that the choice of CWS tools can significantly influence downstream tasks such as sentiment analysis (Lei et al., 2021) and text classification (Zhao et al., 2024). Motivated by this, we hypothesize that CWS tools may also affect topic modeling performance. Building upon the findings from the previous experiment (Table 1, we evaluate the impact of different CWS tools—Jieba, PKUSeg, HanLP, THULAC, and SnowNLP—on topic modeling performance. We chose KmeansTM given its ability to systematically generate coherent and diverse topics. The results (Table S7) indicate that the most notable variations across CWS tools occur in the INT and NPMI metrics. HanLP achieves the highest INT score, followed by Jieba and PKUSeg, while SnowNLP and THULAC perform considerably worse. In terms of NPMI, PKUSeg outperforms all other tools, whereas HanLP reports the lowest score. Since PKUSeg allows the use of domain-specific models, we set the model to a medical domain model for our experimental study. These findings highlight the importance to allow users to chose an appropriate CWS tool, which is of particular relevance for domain-specific datasets.

To further compare the impact of different CWS tools on topic modeling performance, we conducted the Friedman test (Demšar, 2006). The null hypothesis assumes that the choice of CWS tools does not lead to significant differences in topic modeling performance, while the alternative hypothesis posits that such differences do exist. The performance of each tool is measured by multiple topic evaluation metrics, including ISIM, INT, ISH, WESS, EXPRS, NPMI, and COH. The corresponding *p-values* of the Friedman test for these metrics were 4.71×10^{-6}, 1.59×10^{-5}, 3.6×10^{-3}, 0.684, 0.772, 7.58×10^{-4}, and 0.022, respectively. Except for the two metrics of topic diversity, WESS and EXPRS, all *p-values* were below the significance threshold of 0.05, indicating that different CWS tools have a significant impact on topic modeling.

To identify which specific tools differ significantly, we also performed post hoc Nemenyi tests. The results revealed significant differences in ISIM scores between Jieba and HanLP, PKUSeg and HanLP, HanLP and THULAC, and HanLP and SnowNLP. For INT, significant differences were observed between Jieba and THULAC, Jieba and SnowNLP, HanLP and THULAC, and HanLP and SnowNLP. In terms of ISH, HanLP differed significantly from PKUSeg and SnowNLP. For NPMI, PKUSeg and SnowNLP showed significant differences. Finally, for COH, significant differences were found between SnowNLP and Jieba, and SnowNLP and HanLP. Detailed results are presented in Table S8.

Model	hyperparameters
	Dataset: THUCNews_small
LDA	'alpha': 0.251, 'eta': 0.726
NMF	'l1_ratio': 0.042, 'init': 'random', 'max_iter': 742, 'solver': 'mu'
CTM	'encoder_dim': 203, 'dropout': 0.2377, 'inference_type': 'zeroshot', 'lr': 0.0023, 'weight_decay': 0.0010, 'inference_activation': 'Tanh', 'model_type': 'ProdLDA', 'batch_size': 265
BERTopic	'n_neighbors': 31, 'n_components': 11, 'metric': 'euclidean', 'min_cluster_size': 83, 'min_samples': 78, 'cluster_selection_epsilon': 0.698
KmeansTM	'n_neighbors': 20, 'n_components': 30, 'metric': 'euclidean', 'init': 'random', 'n_init': 30, 'max_iter': 293
	Dataset: FUDANCNews
LDA	'alpha': 0.181, 'eta': 0.539
NMF	'l1_ratio': 0.261, 'init': 'random', 'max_iter': 528, 'solver': 'cd'
CTM	'encoder_dim': 511, 'dropout': 0.1805, 'inference_type': 'combined', 'lr': 0.0039, 'weight_decay': 0.0003, 'inference_activation': 'Softplus', 'model_type': 'ProdLDA', 'batch_size': 183
BERTopic	'n_neighbors': 39, 'n_components': 5, 'metric': 'cosine', 'min_cluster_size': 55, 'min_samples': 50, 'cluster_selection_epsilon': 0.288
KmeansTM	'n_neighbors': 21, 'n_components': 37, 'metric': 'euclidean', 'init': 'k-means++', 'n_init': 15, 'max_iter': 992
	Dataset: TOUTIAO_small
LDA	'alpha': 0.271, 'eta': 0.623
NMF	'l1_ratio': 0.746, 'init': 'nndsvda', 'max_iter': 795, 'solver': 'cd'
CTM	'encoder_dim': 308, 'dropout': 0.0046, 'inference_type': 'combined', 'lr': 0.0083, 'weight_decay': 0.0004, 'inference_activation': 'Softplus', 'model_type': 'ProdLDA', 'batch_size': 193
BERTopic	'n_neighbors': 11, 'n_components': 36, 'metric': 'euclidean', 'min_cluster_size': 45, 'min_samples': 99, 'cluster_selection_epsilon': 0.0764
KmeansTM	'n_neighbors': 44, 'n_components': 10, 'metric': 'euclidean', 'init': 'k-means++', 'n_init': 27, 'max_iter': 642
	Dataset: CMtMedQA_small
LDA	'alpha': 0.685, 'eta': 0.228
NMF	'l1_ratio': 0.387, 'init': 'nndsvdar', 'max_iter': 505, 'solver': 'cd'
CTM	'encoder_dim': 134, 'dropout': 0.3819, 'inference_type': 'zeroshot', 'lr': 0.0051, 'weight_decay': 0.0008, 'inference_activation': 'Tanh', 'model_type': 'ProdLDA', 'batch_size': 160
BERTopic	'n_neighbors': 29, 'n_components': 39, 'metric': 'euclidean', 'min_cluster_size': 8, 'min_samples': 76, 'cluster_selection_epsilon': 0.994
KmeansTM	'n_neighbors': 47, 'n_components': 13, 'metric': 'cosine', 'init': 'random', 'n_init': 16, 'max_iter': 662

Table S2: Detailed hyperparameter configurations of each model used in the benchmark experiments reported in Table 1.

THUCNews_small Term Rank	LDA (Topic 5)	NMF (Topic 0)	CTM (Topic 4)	BERTopic (Topic 10)	KmeansTM (Topic 4)
1	游戏 (Game)	游戏 (Game)	游戏 (Game)	游戏 (Game)	游戏 (Game)
2	指数 (Index)	玩家 (Player)	玩家 (Player)	玩家 (Player)	玩家 (Player)
3	玩家 (Player)	海魂 (Haikon)	网游 (Online game)	网游 (Online game)	网游 (Online game)
4	元 (Yuan)	战斗 (Battle)	手机 (Mobile phone)	活动 (Event)	活动 (Event)
5	用户 (User)	角色 (Character)	像素 (Pixel)	装备 (Equipment)	装备 (Equipment)
6	视频 (Video)	体验 (Experience)	体验 (Experience)	奖励 (Reward)	奖励 (Reward)
7	中 (In)	网络游戏 (Online game)	战斗 (Battle)	技能 (Skill)	技能 (Skill)
8	网游 (Online game)	系统 (System)	装备 (Equipment)	战斗 (Battle)	战斗 (Battle)
9	万 (Ten thousand)	平台 (Platform)	活动 (Event)	网络游戏 (Online game)	网络游戏 (Online game)
10	腾讯 (Tencent)	作品 (Production)	奖励 (Reward)	体验 (Experience)	体验 (Experience)

Table S3: Qualitative results for the "Game" topic on the THUCNews_small dataset. The table shows the 10 top words generated by five models along with their English translations (corresponding to Table 1).

TOUTIAO_small Term Rank	LDA (Topic 1)	NMF (Topic 6)	CTM (Topic 1)	BERTopic (Topic 13)	KmeansTM (Topic 6)
1	岁 (Age)	房价 (House prices)	房价 (House prices)	逻辑 (Logic)	房价 (House prices)
2	城市 (City)	买房 (Home purchase)	买房 (Home purchase)	判断 (Judgment)	买房 (Home purchase)
3	买房 (Home purchase)	楼市 (Housing market)	房地产 (Real estate)	马云 (Jack Ma)	房地产 (Real estate)
4	房价 (House prices)	城市 (City)	楼市 (Housing market)	便宜 (Cheap)	楼市 (Housing market)
5	好 (Good)	房地产 (Real estate)	房子 (House)	白菜 (Cabbage)	二手房 (Second-hand housing)
6	地方 (Place)	二手房 (Second-hand housing)	二手房 (Second-hand housing)	未来 (Future)	公积金 (Housing provident fund)
7	最 (Most)	未来 (Future)	城市 (City)	房价 (House prices)	房产 (Property)
8	吃 (Eating)	房产 (Property)	住房 (Housing)	买得起 (Affordable)	住房 (Housing)
9	老婆 (Wife)	调控 (Regulation)	公积金 (Housing provident fund)	房地产 (Real estate)	房子 (House)
10	生活 (Living)	三四 (Third- and fourth-tier)	房产 (Property)	越来越 (Increasingly)	城市 (City)

Table S5: Qualitative results for the "Housing" topic on the TOUTIAO_small dataset. The table shows the 10 top words generated by five models along with their English translations (corresponding to Table 1).

FUDANCNews Term Rank	LDA (Topic 2)	NMF (Topic 10)	CTM (Topic 8)	BERTopic (Topic 12)	KmeansTM (Topic 8)
1	农业 (Agriculture)	农产品 (Agricultural products)	农业 (Agriculture)	种子 (Seed)	农产品 (Agricultural products)
2	生产 (Production)	农民 (Farmers)	发展 (Development)	育种 (Breeding)	粮食 (Grain)
3	粮食 (Grain)	农村 (Rural areas)	农村 (Rural areas)	种子公司 (Seed company)	土地 (Land)
4	发展 (Development)	土地 (Land)	生产 (Production)	科研 (Scientific research)	农户 (Farm households)
5	玉米 (Corn)	产业化 (Industrialization)	农民 (Farmers)	韩国 (South Korea)	产业化 (Industrialization)
6	年 (Year)	农户 (Farm households)	我国 (Our country)	蔬菜 (Vegetables)	农民 (Farmers)
7	小麦 (Wheat)	粮食 (Grain)	农产品 (Agricultural products)	品种 (Cultivar)	农村 (Rural areas)
8	品种 (Cultivar)	经营 (Management)	土地 (Land)	产业 (Industry)	农场 (Farm)
9	提高 (Improvement)	价格 (Price)	市场 (Market)	执法 (Law enforcement)	耕地 (Arable land)
10	经济 (Economy)	科技 (Science and technology)	农户 (Farm households)	专业化 (Specialization)	推广 (Extension)

Table S4: Qualitative results for the "Agriculture" topic on the FUDANCNews dataset. The table shows the 10 top words generated by five models along with their English translations (corresponding to Table 1).

CMtMedQA_small Term Rank	LDA (Topic 9)	NMF (Topic 0)	CTM (Topic 9)	BERTopic (Topic 0)	KmeansTM (Topic 8)
1	症状 (Symptom)	肿瘤 (Tumor)	治疗 (Treatment)	肿瘤 (Tumor)	肿瘤 (Tumor)
2	治疗 (Treatment)	靶向 (Targeted therapy)	肿瘤 (Tumor)	术后 (Postoperative)	化疗 (Chemotherapy)
3	医生 (Doctor)	化疗 (Chemotherapy)	手术 (Surgery)	牙齿 (Tooth)	放疗 (Radiotherapy)
4	肺癌 (Lung cancer)	放疗 (Radiotherapy)	化疗 (Chemotherapy)	评估 (Assessment)	切除 (Resection)
5	缓解 (Relief)	治疗 (Treatment)	放疗 (Radiotherapy)	子宫 (Uterus)	肺癌 (Lung cancer)
6	情况 (Condition)	癌细胞 (Cancer cell)	患者 (Patient)	这种 (This)	甲状腺 (Thyroid)
7	药物 (Drug)	细胞 (Cell)	切除 (Resection)	具体情况 (Specific condition)	靶向 (Targeted therapy)
8	进行 (Carry out)	切除 (Resection)	需要 (Need)	心理 (Psychological)	乳腺 (Mammary glands)
9	建议 (Advice)	副作用 (Side effect)	靶向 (Targeted therapy)	化疗 (Chemotherapy)	乳腺癌 (Mammary cancer)
10	引起 (Cause)	方案 (Protocol)	肺癌 (Lung cancer)	考虑 (Consideration)	乳房 (Breast)

Table S6: Qualitative results for the "Oncology" topic on the CMtMedQA_small dataset. The table shows the 10 top words generated by five models along with their English translations (corresponding to Table 1).

metrics	Intruder			Diversity		Coherence	
CWS tool	ISIM ↓	INT ↑	ISH ↓	WESS ↓	EXPRS ↓	NPMI ↑	COH ↑
Jieba	0.619 (0.004)	0.573 (0.064)	0.707 (0.003)	0.798 (0.007)	**0.818** (0.005)	0.271 (0.031)	0.744 (0.010)
PKUSeg	**0.616** (0.004)	0.517 (0.070)	**0.706** (0.003)	**0.797** (0.007)	**0.818** (0.006)	**0.290** (0.031)	0.740 (0.009)
HanLP	0.623 (0.004)	**0.576** (0.059)	0.711 (0.004)	0.802 (0.009)	**0.818** (0.006)	0.254 (0.038)	**0.746** (0.008)
THULAC	0.618 (0.002)	0.499 (0.042)	0.708 (0.003)	0.800 (0.006)	0.819 (0.003)	0.260 (0.028)	0.738 (0.005)
SnowNLP	0.616 (0.004)	0.511 (0.054)	**0.706** (0.004)	0.800 (0.007)	0.819 (0.004)	0.256 (0.028)	0.735 (0.006)

Table S7: Effects of CWS tool on the results of KmeansTM topic model applied to the CMtMedQA_small dataset. In the preprocessing stage for the CMtMedQA_small dataset, we use varying CWS tools (Jieba, PKUSeg, HanLP, THULAC, and SnowNLP) while keeping the other processes unchanged. For each dataset processed by a different CWS tool, we compute key intruder, diversity, and coherence metrics based on the results of 20 runs of KmeansTM topic model with mean and standard deviation (in parentheses). The ↑ symbol indicates that a higher value is better, while the ↓ symbol indicates that a lower value is better.

	jieba	pkuseg	hanlp	thulac	snownlp
			ISIM		
jieba	-	0.497	**0.041**	1.000	0.180
pkuseg		-	**0.000**	0.497	0.975
hanlp			-	**0.041**	**0.000**
thulac				-	0.180
snownlp					-
			INT		
jieba	-	0.562	0.807	**0.009**	**0.041**
pkuseg		-	0.070	0.373	0.691
hanlp			-	**0.000**	**0.000**
thulac				-	0.987
snownlp					-
			ISH		
jieba	-	0.975	0.054	0.497	1.000
pkuseg		-	**0.009**	0.180	0.995
hanlp			-	0.807	**0.031**
thulac				-	0.373
snownlp					-
			NPMI		
jieba	-	0.562	0.807	0.897	0.562
pkuseg		-	0.070	0.115	**0.022**
hanlp			-	1.000	0.995
thulac				-	0.975
snownlp					-
			COH		
jieba	-	0.931	0.751	0.562	**0.031**
pkuseg		-	0.266	0.957	0.220
hanlp			-	0.054	**0.000**
thulac				-	0.628
snownlp					-

Table S8: Nemenyi Test Results (*p-values*) for ISIM, INT, ISH, NPMI and COH. *p-values* less than 0.05 are in bold.

Dataset	Total	Category Distribution (Class: Documents count)
THUCNews_imbalance	8,400	Finance: 3000, Sports: 3000, Entertainment: 800, Technology: 800, Education: 200, Real Estate: 200, Fashion: 200, Politics: 200, Home: 200, Stock: 200, Society: 200, Constellation: 200, Game: 200, Lottery: 200.
FUDANCNews	9,526	Sports: 1461, Computer: 1141, Economy: 1139, Agriculture: 1092, Politics: 1090, Environment: 959, Space: 890, Art: 537, History: 243, Military: 146, Education: 116, Transport: 116, Medical: 104, Law: 102, Philosophy: 85, Literature: 67, Mining: 67, Energy: 65, Electronics: 55, Communication: 51.
CMtMedQA_imbalance	10,000	Pediatrics: 3000, Surgery: 3000, OB/GYN: 1000, ENT: 1000, Internal Med: 1000, Dermatology: 200, Infectious Dis: 200, Oncology: 200, Psychology: 200, Andrology: 200.
TOUTIAO_imbalance	19,600	Agriculture: 3000, Tech: 3000, Entertainment: 2200, Sports: 2000, Finance: 1600, Culture: 1400, Education: 1200, Housing: 1100, Game: 900, Travel: 900, Automotive: 800, Military: 800, World: 500, Story: 200.

Table S9: Category distribution of the four imbalance datasets.

Metrics	Intruder			Diversity		Coherence	
Model	ISIM ↓	INT ↑	ISH ↓	WESS ↓	EXPRS ↓	NPMI ↑	COH ↑
Dataset: THUCNews_imbalance							
LDA	0.572	0.175	0.717	0.885	0.896	-0.612	0.597
NMF	0.533	0.403	0.644	0.773	0.843	0.394	0.644
CTM	0.540	0.337	0.654	0.785	0.835	0.321	0.633
BERTopic	0.557	**0.485**	0.657	0.762	0.843	**0.412**	**0.686**
KmeansTM	**0.508**	0.472	**0.617**	**0.735**	**0.826**	0.377	0.658
Dataset: TOUTIAO_imbalance							
LDA	0.511	0.146	0.668	0.845	0.886	-0.601	0.546
NMF	0.526	0.267	0.657	0.809	0.870	**-0.067**	0.605
CTM	0.501	0.522	0.600	0.704	0.810	-0.120	**0.668**
BERTopic	0.606	0.209	0.745	0.895	0.919	-0.436	0.628
KmeansTM	**0.481**	**0.560**	**0.584**	**0.691**	**0.804**	-0.096	0.647
Dataset: CMtMedQA_imbalance							
LDA	0.639	0.123	0.782	0.942	0.913	-0.089	0.636
NMF	0.614	0.164	0.745	0.886	0.883	0.213	0.658
CTM	**0.609**	**0.574**	**0.703**	0.810	0.849	0.110	0.715
BERTopic	0.642	0.290	0.749	0.859	0.864	**0.236**	0.707
KmeansTM	0.621	0.444	0.716	**0.803**	**0.818**	**0.236**	**0.730**

Table S10: Sensitivity results on the imbalance Chinese datasets. All models and evaluation metrics use the Conan-embedding-v1 pre-trained embedding model (Li et al., 2024) when applicable. Model hyperparameters are tuned over 20 trials. The intruder metrics are averaged over 100 runs. All models, except BERTopic, are fitted with a pre-specified number of 14, 20, 14, and 10 topics, respectively. BERTopic automatically detects the optimal number of topics, hence we extract the corresponding number of topics from the output in order. The ↑ symbol indicates that a higher value is better, while the ↓ symbol indicates that a lower value is better.

Metrics	Intruder			Diversity		Coherence	
n_topics (K)	ISIM ↓	INT ↑	ISH ↓	WESS ↓	EXPRS ↓	NPMI ↑	COH ↑
Dataset: THUCNews_small							
10	**0.483**	**0.572**	**0.589**	**0.707**	**0.817**	**0.457**	0.634
30	0.500	0.541	0.604	0.726	0.825	0.447	0.651
50	0.501	0.513	0.606	0.721	0.820	0.399	0.650
100	0.516	0.509	0.622	0.743	0.832	0.360	**0.658**

Table S11: Sensitivity results on the THUCNews_small dataset. KmeansTM and evaluation metrics use the Conan-embedding-v1 pre-trained embedding model (Li et al., 2024) when applicable. Model hyperparameters are tuned over 20 trials. The intruder metrics are averaged over 100 runs. KmeansTM is fitted with a increasing number of topics (K). The ↑ symbol indicates that a higher value is better, while the ↓ symbol indicates that a lower value is better.