**ARTICLE**

# GLMTopic: A Hybrid Chinese Topic Model Leveraging Large Language Models

**Weisi Chen**[1,*], **Walayat Hussain**[2,*] **and Junjie Chen**[1]

[1]School of Software Engineering, Xiamen University of Technology, Xiamen, 361024, China
[2]Peter Faber Business School, Australian Catholic University, North Sydney, 2060, Australia
*Corresponding Authors: Weisi Chen. Email: chenweisi@xmut.edu.cn; Walayat Hussain. Email: walayat.hussain@acu.edu.au

**ABSTRACT:** Topic modeling is a fundamental technique of content analysis in natural language processing, widely applied in domains such as social sciences and finance. In the era of digital communication, social scientists increasingly rely on large-scale social media data to explore public discourse, collective behavior, and emerging social concerns. However, traditional models like Latent Dirichlet Allocation (LDA) and neural topic models like BERTopic struggle to capture deep semantic structures in short-text datasets, especially in complex non-English languages like Chinese. This paper presents Generative Language Model Topic (GLMTopic) a novel hybrid topic modeling framework leveraging the capabilities of large language models, designed to support social science research by uncovering coherent and interpretable themes from Chinese social media platforms. GLMTopic integrates Adaptive Community-enhanced Graph Embedding for advanced semantic representation, Uniform Manifold Approximation and Projection-based (UMAP-based) dimensionality reduction, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering, and large language model-powered (LLM-powered) representation tuning to generate more contextually relevant and interpretable topics. By reducing dependence on extensive text preprocessing and human expert intervention in post-analysis topic label annotation, GLMTopic facilitates a fully automated and user-friendly topic extraction process. Experimental evaluations on a social media dataset sourced from Weibo demonstrate that GLMTopic outperforms Latent Dirichlet Allocation (LDA) and BERTopic in coherence score and usability with automated interpretation, providing a more scalable and semantically accurate solution for Chinese topic modeling. Future research will explore optimizing computational efficiency, integrating knowledge graphs and sentiment analysis for more complicated workflows, and extending the framework for real-time and multilingual topic modeling.

**KEYWORDS:** Topic modeling; large language model; deep learning; natural language processing; text mining

## 1 Introduction

### 1.1 Background and Motivation

Topic modeling (TM) is an unsupervised machine learning method for extracting clustered topics or themes from a corpus of documents, such as books, websites, blogs, social media posts, emails, news, or research articles etc. TM has long been useful in natural language processing (NLP) [1].

In line with the advancement of artificial intelligence, NLP techniques, especially in the field of TM, have been maturing from rule-based, statistical methods to machine learning and deep learning ones. Over the last decade, mainstream TM techniques such as Latent Dirichlet Allocation (LDA) [2], Top2Vec [3], and BERTopic [4] have gained widespread adoption. Specifically, LDA, a traditional statistical model based on word frequency and co-occurrence relationships, has been the most predominant model for a long time and

continues to influence various domains. As deep learning has evolved in recent years, neural network-based models like BERTopic have gained popularity for their ability to explore potential topics more deeply, thanks to advanced deep learning frameworks.

The motivation behind this study stems from the limitations of traditional topic modeling techniques when applied to Chinese short-text datasets. Despite the significant progress of TM techniques, existing models still struggle to capture complex linguistic patterns and subtle differences in tone within the text, especially in complex languages like Chinese, resulting in the extracted topics falling short of coherence and consistency. With the recent rise of large language models (LLMs), there is an opportunity to enhance topic modeling by leveraging their contextual understanding and reasoning abilities. This paper aims to develop a model that integrates LLMs with appropriate techniques to improve topic coherence. Additionally, the study seeks to address the challenge of automating topic interpretation, reducing the reliance on human experts for labeling and refining results. The objective is to create a scalable, efficient, and semantically rich topic modeling solution that is particularly effective for Chinese short-text data, making it more accessible for researchers, businesses, and policymakers.

### 1.2 Research Questions and Contributions

This paper attempts to address the following research questions (RQs):

**RQ 1: Can we leverage the capabilities of LLMs to improve Chinese topic modeling performance for social science research on large-scale social media data?**

This research question explores the feasibility of developing a novel hybrid topic modeling approach that integrates LLMs to enhance topic coherence and semantic understanding, particularly in Chinese short-text datasets. Current models, such as LDA and BERTopic, face challenges in capturing complex linguistic patterns and maintaining semantic consistency, especially in non-English languages like Chinese. This paper aims to address these limitations and develop a more effective topic modeling framework to facilitate social science researchers in identifying and understanding key themes discussed across various Chinese social media platforms.

**RQ 2: How can we interpret the topic modeling results without the intervention of human experts to enhance applicability?**

This research question focuses on improving the interpretability of topic modeling results by minimizing the reliance on human domain experts for post-processing and validation. Many traditional topic modeling methods require significant manual effort to refine, label, and make sense of extracted topics. By integrating LLMs with an advanced topic modeling framework, the study seeks to automate the interpretation process, ensuring that the generated topics are more coherent, directly meaningful, and accessible to users across different domains without requiring specialized expertise.

To address these RQs, this paper proposes a novel LLM-empowered hybrid model named GLMTopic. This model takes advantage of the capabilities of the most cutting-edge technologies to capture deep semantic and contextual information and enhance the coherence of topic discovery, especially in the domain of short Chinese texts. GLMTopic employs Adaptive Community-enhanced Graph Embedding (ACGE), which incorporates Matryoshka Representation Learning (MRL) to generate nested, multi-granular semantic embeddings. MRL produces a hierarchically structured representation where each sub-vector captures semantic information at different levels of granularity, allowing for flexible adaptation to various computational resources and downstream tasks. This enables GLMTopic to maintain both semantic richness and representational efficiency. Moreover, ACGE's integration of graph-based community enhancement facilitates the construction of high-level semantic relations across short texts, which is especially beneficial

in modeling the fragmented and context-rich discourse found on social media platforms. This approach supports the identification of socially relevant themes and underlying group concerns, providing valuable insights for social science researchers studying public discourse and collective behavior.

The rest of the paper is structured as follows: Section 2 discusses recent progress on topic modeling techniques. Section 3 details the proposed hybrid TM model leveraging an LLM named GLMTopic, describing its design principles and how it overcomes the limitations of traditional TM techniques. Section 4 presents an experiment evaluating the proposed model against conventional methods, including LDA and BERTopic. Section 5 summarizes the paper and discusses future research directions.

## 2 Related Work

From the late 1980s to the early 1990s, TM technology was initially dominated by the bag-of-words model (BoW) and the word frequency-inverse document frequency (TF-IDF) based on word frequency statistics. Later, LDA, a probabilistic generative model, was introduced in 2003 [2], using a Dirichlet prior distribution for document-topic and topic-word distributions, significantly improving the topic extraction capability. It has been widely employed since its inception in various domains such as literature review for research [5], news analysis [6], social science [7], etc. However, traditional methods like LDA often require extensive corpus preprocessing, careful parameter tuning, thorough model evaluation, and domain expertise to interpret the generated topics effectively. Neural modeling has evolved rapidly since 2016 with the evolving capabilities of artificial neural networks and deep learning. These advanced methods further improve topic generation's accuracy and semantic capture ability. The representative neural topic models include DeepLDA, Top2Vec, and BERTopic. Specifically, DeepLDA is a hybrid model that combines LDA with basic multilayer perceptron (MLP) neural networks [8]. At the same time, the more cutting-edge BERTopic utilizes advanced techniques such as bi-directional encoder representation from transformers (BERT) and class-based TF-IDF (c-TF-IDF). More details about these models can also be found in a recent study [9].

BERTopic has gained prominence in recent years because it produces reliable results with minimal data preprocessing on various datasets. Recent literature has reported its superiority to some other models. As shown in Table 1, the application of BERTopic has grown rapidly, especially after 2022. This demonstrates its effectiveness across various domains and datasets, promoting academic research and practical applications.

**Table 1:** Representative applied research using topic modeling in recent literature

| Ref. | Year | Topic model | Description |
|------|------|-------------|-------------|
| [10] | 2025 | LDA BERTopic NMF | Analyzing GoPartner reviews to identify driver-related issues in Indonesia's transportation and food delivery sector. BERTopic and LDA produced higher-quality topics. |
| [11] | 2024 | BERTopic | Analyzing COVID-related news from New York Times to understand the evolution of media narratives and sentiment during and after the COVID pandemic using BERTopic for topic modeling and RoBERTa for sentiment analysis. Identified shifting public concerns from outbreak panic to long-term societal and economic impacts. |
| [12] | 2024 | BERTopic | Analyzing the data of "People's Letters" in Zhejiang Province. The core themes of public demand are concentrated in the aspects of life security and rights protection, with special attention to the response to emergencies. |

(Continued)

**Table 1 (continued)**

| Ref. | Year | Topic model | Description |
|:---:|:---:|:---:|:---:|
| [13] | 2024 | BERTopic | A combination of BERTopic and social network analysis was used to identify the main topics about diabetes among Indian Twitter users and their path of information dissemination. |
| [14] | 2024 | BERT + hyperparameter optimization | Exploring the best hyperparameter tuning methods for topic modeling in the BERT model, identify the key hyperparameters that affect the performance of BERT topic modeling, and provide optimization recommendations. |
| [15] | 2024 | BERTopic | Analyzing the literature related to grassland degradation and to reveal the thematic structure and trends in the research field. |
| [16] | 2024 | LDA Top2Vec BERTopic | Analyzing 55,442 library science papers, revealing differences in topic granularity and document distribution across models. |
| [17] | 2024 | BERTopic | Analyzing 3767 AI research articles, identifying 19 clusters, including AI in health, sustainability, and ethics. BERTopic highlights key trends and interconnected themes, showcasing its value in mapping AI research advancements. |
| [18] | 2024 | BERTopic | Analyzing stock market comments, integrating topic sentiment with deep learning models to improve stock price prediction and reveal market insights. |
| [19] | 2024 | LDA | A systematic review of the intersection between machine learning and mobile crowdsourcing. It applied LDA to extract key research topics from 133 research papers, revealing trends such as federated learning, data labeling, task allocation, and privacy concerns in mobile crowdsourcing systems. |
| [9] | 2023 | LDA Top2Vec BERTopic | Analyzing long financial news, compares the performance of these models, demonstrating BERTopic's superior coherence. The paper also introduces the News Impact Analysis (NIA) framework, which integrates multiple topic models to automate financial market impact analysis. |
| [20] | 2023 | LDA LSA Top2Vec | Compareing topic models on 65,292 COVID-19 abstracts, highlighting strong topic alignment between Top2Vec and LDA. Practical recommendations are provided. |
| [21] | 2023 | LDA | Analyzing 177,204 AI research documents (1990–2021), identifying 7 subfields and revealing topic distributions by year, source, and region to guide research directions. |
| [22] | 2023 | LDA | Analyzing 33,957 articles in fuzzy research, identifying 10 key topics, their trends, and distribution across journals and regions, offering insights for future research and collaboration. |
| [23] | 2023 | LDA BERTopic | Analyzing public perceptions of climate change using Twitter data. BERTopic achieved high performance with 91.35% precision, 89.65% recall, and 93.50% accuracy. |

Although the traditional TM technique LDA is still frequently used, LDA but its core mechanism relies on the bag-of-words model (BoW), which is only based on word frequency information to derive topic distribution, its cumbersome preprocessing processes, such as word splitting and deactivation as well as based on the cleanup of a specific domain-specialized text tend to add to the burden of researchers. The emergence of BERTopic has brought new vitality to the field of topic modeling, with its capability to capture longer temporal dependencies within the text. However, BERTopic still has limitations when dealing with extremely complex contexts or highly domain-specific corpora. Also, most existing models rely on human experts to annotate the topic labels based on the keywords of each topic, which is a non-trivial task.

The trending LLM technologies like GPT [24] and GLM [25], with their superior contextual understanding and deep semantic understanding of text, have been introduced to break through the limitations of traditional NLP techniques, with their comprehensive NLP capabilities gained from massive training using a large amount of textual data. LLM [26,27] has been extensively applied on various fronts, which has inspired us to explore the possibility of leveraging it in topic modeling, aiming to resolve the remaining problems among existing TM techniques. The introduction of LLMs can potentially break through traditional TM technology's limitations and reduce the dependence on complex text preprocessing and domain knowledge, thus significantly improving topic coherence and reducing the user's operational burden. In addition to LLMs, Visual Language Models (VLMs) have also shown their great potential in multimodal data processing. Studies have shown that VLMs perform significantly better in cross-modal tasks, such as multimodal news analysis, where the combination of images and text provides richer contextual information for understanding news content [28,29].

## 3 GLMTopic: An LLM-Based Topic Model

While LDA and BERTopic are currently among the most dominant TM models, they do not systematically specify a standard process for topic modeling and serve more as tools than systematic processes. Therefore, we propose a hybrid model, GLMTopic, which aims to leverage the powerful capabilities of LLMs for topic modeling. This section will introduce the structure and constituents of GLMTopic.

### 3.1 Model Structure

This study proposes a hybrid model with a large language model at its core, named GLMTopic. Compared with traditional topic modeling methods, GLMTopic integrates the advantages of statistical methods and open-source large language models and achieves deep integration and optimization of functional modules in various aspects, such as text semantic embedding, dimensionality reduction, clustering, keyword extraction, and semantic consistency optimization.

Specifically, as shown in Fig. 1, the original documents are processed using a word embedding model named Adaptive Community-enhanced Graph Embedding (ACGE) and turned into vectors, denoted as $V_{1,1}$, $V_{1,2,...}$, $V_{1,n,...}$, where $V_{i,j}$ represents the word vector of the j-th word in the i-th document. These word vectors are then downscaled using UMAP to reduce dimensionality and obtain the lower-dimensional representations, denoted as $V_1$, $V_2$,..., $V_n$, which are subsequently clustered using HDBSCAN to identify potential thematic structures. Next, CountVectorizer and c-TF-IDF are used to generate a word frequency matrix for the text in each cluster, to determine the most important keywords for each topic. The topic keywords generated by c-TF-IDF, as well as the original documents, are fed into an LLM named GLM to find the most appropriate topic representations, further improving the semantic consistency. Finally, the topic representations are turned into the visualization of the results. More details about each component of this hybrid model will be presented in the following section.
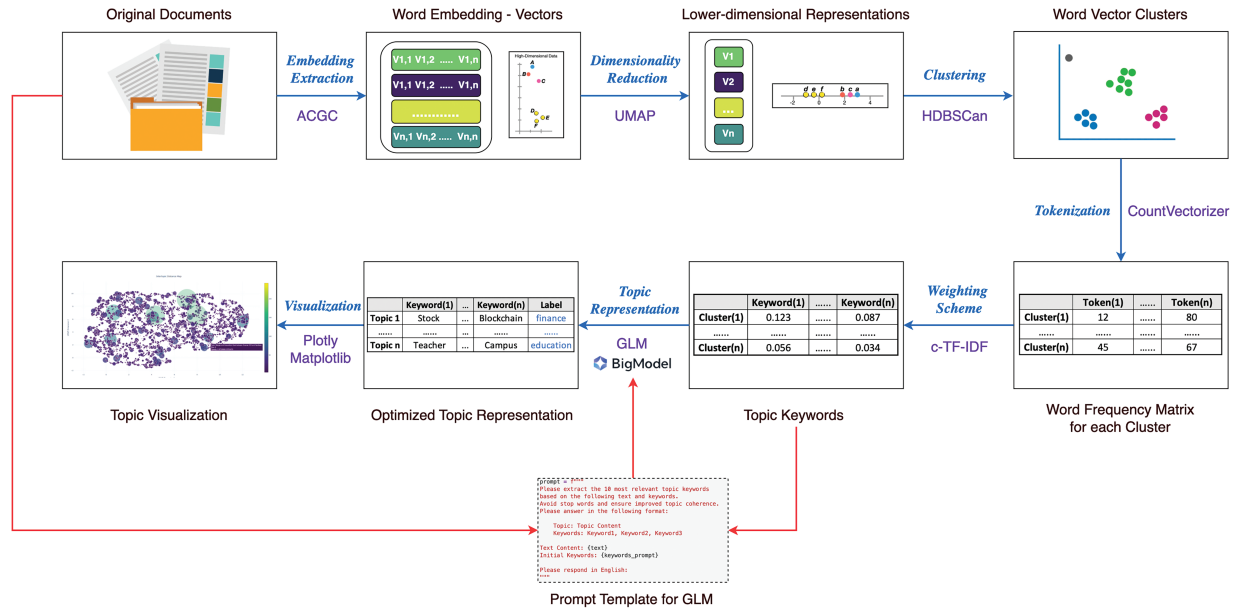
**Figure 1:** Structure of the GLMTopic hybrid model

## 3.2 Model Constituents

### 3.2.1 Embedding

The GLMTopic hybrid model adopts ACGE (Adaptive Community-enhanced Graph Embedding: https://huggingface.co/aspire/acge_text_embedding (accessed on 03 July 2025).), which combines community enhancement and graph neural network techniques to improve the accuracy of semantic modeling significantly. Compared with a mainstream Sentence-BERT (SBERT) model [30], the ACGE model can better capture the deep implicit semantic associations in the text by introducing community structure features, which are especially suitable for processing the results of polysemous and complex syntax in noontime text. While the embedding of SBERT relies on the semantic embedding of dominant features between sentences, ACGE can extract more comprehensive semantic relations through the community detection technique and the graph neural network. As a result, ACGE shows stronger adaptability and modeling ability in processing long texts (e.g., news reports and research papers) and short texts (e.g., Tweets and Weibo posts).

It is worth noting that ACGE provides GLMTopic with the advantage of requiring minimal preprocessing. Traditional topic modeling requires cumbersome text preprocessing, such as deletion of stop words, word splitting, noise processing, etc. In contrast, ACGE combines graph embeddings and community-adding effects to directly process preprocessed text data, especially for complex expressions in Chinese contexts. For complex expressions in Chinese contexts, ACGE automatically extracts key semantics through community structure, reducing the need for a preprocessing process and improving the GLMTopic model's ability to retain the original text's semantic information.

### 3.2.2 Dimensionality Reduction and Clustering

In the dimensionality reduction stage, the GLMTopic model employs Uniform Manifold Approximation and Projection (UMAP) [31] to reduce the dimensionality of the high-dimensional vector embeddings. UMAP ensures the integrity of the semantic information while reducing the computational complexity. UMAP can better preserve the local and global structural features in the high-dimensional space, providing

better quality basic data for the later clustering task. After dimensionality reduction, the GLMTopic model adopts the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering algorithm [32]. Considering that the number of topics embedded in a large amount of textual data is still unknown, the HDBSCAN can be adapted by the HDBSCAN can automatically determine the optimal number of clusters according to the semantic similarity and group similar texts into one class in an adaptive way, thus avoiding the problem of setting the number of clusters in traditional algorithms such as K-means. Therefore, it is especially suitable for dealing with unknown and uneven subject distribution of text data. In addition, the robustness of HDBSCAN to noise significantly improves the accuracy of clustering results.

### 3.2.3 Word Splitting and Weight Optimization

GLMTopic utilizes CountVectorizer as a word splitter, which splits the text into tokens and counts the frequencies to generate a word frequency matrix. CountVectorizer can split the text into words or n-grams while retaining the original underlying statistical information. On this basis, GLMTopic uses the c-TD-IDF (cluster-based Term Frequency-Inverse Document Frequency) algorithm, which combines the inter-topic information based on CountVectorizer's word frequency matrix and optimizes the weights of topic keywords. Optimized. Compared with the original TF-IDF, it can be more applicable to the semantic distribution of clustering results and can be better cultivated for topic modeling scenarios. Through the weighted optimization of c-TD-IDF, the GLMTopic model can more accurately extract the theme core vocabulary and significantly enhance the reasonableness of the semantic distribution. The formula is as follows:

$$c - TD - IDF\left(t, c\right) = f\left(t, c\right) \times \log\left(\frac{N}{n_e}\right) \tag{1}$$

where $f(t,c)$ denotes the term frequency of word $t$ in class $c$, i.e., how many times the word $t$ appears in class $c$; $N$ is the total number of documents in the corpus; and $n_e$ represents the number of documents in class $c$ that contain the term $t$.

### 3.2.4 Representation Tuning Using GLM

For semantic consistency optimization, GLMTopic jointly inputs the topic keywords and their weights extracted by c-TD-IDF with the original content into GLM for language verification and optimization, and GLM, using its powerful contextual understanding capability, can perform correlation analysis between the topic keywords and the original content to validate the reasonableness of the topic keywords and correct the possible redundancy and semantic bias. At the same time, GLM dynamically adjusts the list of topic terms by matching the topic keywords with the context to make it closer to the core semantics of the original text, thus reaching the optimized topic coherence and consistency. The GLMTopic model realizes the profound fusion of the statistical method and the big language model, which improves the semantic accuracy and the class interpretability of the topic modeling. This process makes the extracted topics of GLMTopic more closely matched to the core semantics of the original text when processing complex textual fields.

Specifically, in this study we adopt GLM-4-Long[1] in the GLM family, an open-source LLM developed by Zhipu AI, known for its strong capabilities in contextual understanding of Chinese texts. The LLM is used exclusively in this final representation tuning phase to refine the keywords associated with each topic and ensure better alignment between the topic representation and the underlying document content. For each discovered topic, we extract the top-n keywords using class-based TF-IDF (c-TF-IDF), and then pass these

---

[1] https://huggingface.co/THUDM/glm-4-9b-chat-1m (accessed on 03 July 2025).

keywords, along with representative documents from the cluster, to the LLM. This step is conducted in a zero-shot setting, without further fine-tuning. GLM-4-Long is deployed locally, and its integration helps address common weaknesses in topic modeling such as redundancy, vague keyword sets, or disconnected terms.

### 3.2.5 Topic Visualization

To display the modeling results more intuitively, GLMTopic adopts the Plotly library to visualize the output of the topic modeling results and generates an interactive topic-word relationship graph. The relationship graph shows the semantic associations among the topic words. It helps users explore the topic structure and semantic distribution in depth through interactive operations. The interactive features further enhance the results' interpretability and analytic value and help discover the hidden connections between topics.

To sum up, the GLMTopic model, through modular design and in-depth optimization, gives full play to the powerful capabilities of large language models in semantic understanding and complex language processing and provides an efficient solution for multi-scenario topic modeling tasks.

## 4 Experiments and Results

### 4.1 Dataset

The dataset in this experiment has been sourced from Weibo, a mainstream Chinese microblogging platform, with a total of 22,184 posts, totaling 10.8 MB. These posts were collected from multiple categories of Weibo Hot Searches, including international, real-time, inner-city, e-sports, celebrity, fun, emotion, etc. The average number of Chinese characters in each Weibo post is 61.44. Figs. 2 and 3 illustrate Chinese and English samples from the dataset, respectively.

| | content | type | create_at | authorName |
|---|---|---|---|---|
| 0 | 【俄乌进入谈判前"至暗时刻"？泽连斯基最新言论已暗含变化？】俄罗斯国防部19日证实，乌克兰当... | 国际 | 2024-11-21 | 凤凰卫视 |
| 1 | 【废除法规 解雇员工 #马斯克公布大规模政府裁员计划#】据央视新闻消息，当地时间11月20日... | 国际 | 2024-11-21 | 贝壳财经 |
| 2 | 11月20日，乌克兰使用英国"风暴阴影"巡航导弹，攻击了俄罗斯境内目标。按照俄罗斯国防部的战... | 国际 | 2024-11-21 | 三叔的装备空间 |
| 3 | 牛市结束，市场回归平静。最终资金还是要选择业绩股，低价股，但是市场上真正的低价股还有多少？新... | 国际 | 2024-11-21 | 磐石之心 |
| 4 | 马斯克新论:美国就是新罗马。#马斯克# #特朗普新任命# | 国际 | 2024-11-21 | 羾枘萨满 |
| ... | ... | ... | ... | ... |
| 22179 | 巴勒斯坦"卡桑旅"（哈马斯）发布新视频，在加沙地区中部走廊狙击一名以色列士兵，使用的是AM-... | 巴以 | 2024-11-16 | 沙姆雄狮_EL |
| 22180 | #巴以冲突##看见具体的巴勒斯坦人# 巴勒斯坦记者Hossam·Shabat和加沙平民近照 | 巴以 | 2024-11-16 | 灰色魔毯Enterprise-NorthPoint |
| 22181 | #巴以冲突# 圣城新闻网：以色列军车和直升机在加沙地带北部贝特拉希耶西部地区猛烈扫射，同时还... | 巴以 | 2024-11-16 | 灰色魔毯Enterprise-NorthPoint |
| 22182 | 11月15日，在德方倡议下，普京与朔尔茨进行了自2022年12月以来的首次电话会谈。双方就乌... | 巴以 | 2024-11-16 | 墨坛 |
| 22183 | 今天一群犹太人暴徒袭击巴勒斯坦拜特福里克镇阿尔杜尔巴特社区的结果，许多原住民的房屋遭到暴徒的纵... | 巴以 | 2024-11-16 | 沙姆雄狮_EL |

22184 rows × 4 columns

**Figure 2:** Sample Weibo data used in the experiment

| | content | type | create_at | authorName |
|---|---|---|---|---|
| 0 | [Are Russia and Ukraine entering the "darkest ... | International | 2024-11-21 | Phoenix TV |
| 1 | [Abolishing regulations, laying off employees ... | International | 2024-11-21 | Shell Finance |
| 2 | On November 20, Ukraine used UK-supplied 'Stor... | International | 2024-11-21 | Uncle San's Gear Space |
| 3 | The bull market has ended and the market is re... | International | 2024-11-21 | Heart of Rock |
| 4 | Musk's new theory: America is the new Rome. #M... | International | 2024-11-21 | Zangke Shaman |
| ... | ... | ... | ... | ... |
| 22179 | The Palestinian 'Al-Qassam Brigades' (Hamas) r... | Israel - Palestine | 2024-11-16 | Lion of Sham_EL |
| 22180 | #IsraelPalestineConflict #SeePalestiniansAsPeo... | Israel - Palestine | 2024-11-16 | Gray Carpet Enterprise-NorthPoint |
| 22181 | #IsraelPalestineConflict# Holy City News: Isra... | Israel - Palestine | 2024-11-16 | Gray Carpet Enterprise-NorthPoint |
| 22182 | On November 15, at Germany's initiative, Putin... | Israel - Palestine | 2024-11-16 | Ink Forum |
| 22183 | Today, a group of Jewish extremists attacked t... | Israel - Palestine | 2024-11-16 | Lion of Sham_EL |

22184 rows × 4 columns

**Figure 3:** Sample Weibo data used in the experiment (English translation)

### 4.2 Hardware and Software Settings

Three topic models, LDA, BERTopic, and GLMTopic, were implemented in Python, using libraries including scikit-learn, BERTopic, and the API of Zhipu AI, respectively. Table 2 details the implementation of these topic models, including the required Python libraries and the parameters used in this study.

**Table 2:** Details of topic model implementation in this study

| Item | LDA | BERTopic | GLMTopic |
|---|---|---|---|
| Python library | Scikit-learn | Bertopic | Zhipu AI |
| Version | 1.5.2 | 0.16.3 | 2.1.5.20230904 |
| No. of topics | 50, 100, 150, 200, 250, 300, 350, 400 | Undefined | Undefined |
| Embedding Extraction | Topic-word and document-topic distributions | SBERT | ACGE |
| Dimensionality reduction | – | UMAP | UMAP |
| Clustering | Probability-based | HDBSCAN | HDBSCAN |
| Weight scheme | TF-IDF and Dirichlet priors | c-TF-IDF | c-TF-IDF |
| Representation tuning | Default | MMR | GLM |

The hardware settings where the experiments were conducted are as follows.

- Operating System: MacOS Sonoma 14.2;
- CPU: Apple M1 8 GB; RAM: 8 cores.

### 4.3 Results

#### 4.3.1 LDA Results

We trained the LDA model using eight different sets of parameters, with parameter combinations including various numbers of topics (50, 100, 150, 200, 250, 300, 350, 400) and other hyperparameters under default settings. We recorded the coherence score of all the topics generated in each run. These topics and keywords were analyzed in detail, and an attempt was made to interpret the results and provide topic annotations. These annotations helped us understand the potential associations between different themes

and industry sectors. Tables 3 and 4 show the top 10 examples of themes generated by the LDA-150 (150 topics) and LDA-300 (300 topics) models, respectively. Tables 5 and 6 present the English versions of Tables 3 and 4, respectively. Some of them have keywords that have little significance for social researchers and have been marked as blank in the tables. Fig. 4 shows the results of the topic visualization generated by the LDA-150 model, reflecting the statistical proximity between the different topics.

**Table 3:** Sample topics generated by LDA 150 (150 topics—Chinese version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | 邀请, 更好, 举办, 大赛, 之下, 朋友圈, 整体, 大师, 中式, 微博美妆 | 美妆活动 |
| Topic 2 | 未来, 捐赠, 产业, 队伍, 农业, 固态, 电池, 反击, 项目, 培养 | 产业发展 |
| Topic 3 | 内容, 周六, 唯一, 阅读, 远离, 话题, 气质, 听说, 东北, 节目 | – |
| Topic 4 | 依然, 市民, 驻足, 多年, 内阁, 公共, 当日, 建成, 呈现, 广场 | 城市建设 |
| Topic 5 | 研究, 高级, 完整, 导演, 协商, 鼓掌, 一生, 看过, 神话, 两部 | 电影 |
| Topic 6 | 保护, 无人机, 攻击, 多地, 司机, 政府, 权利, 一项, 一群, 交易 | 安全与政策 |
| Topic 7 | 暂时, 战时, 手册, 关心, 开放, 世纪, 休息, 一根, 吃惊, 采访 | 战时反应 |
| Topic 8 | 告诉, 估计, 风险, 人数, 改善, 否认, 集团, 创造, 一道, 咨询 | 风险管理 |
| Topic 9 | 秘鲁, 手术, 专家, 品牌, 偷笑, 大型, 两位, 嫌疑人, 致敬, 一手 | 医疗 |
| Topic 10 | 城市, 办公室, 守护, 冬季, 福建, 三年, 观察, 节奏, 一线, 颜值 | 城市 |

**Table 4:** Sample topics generated by LDA 300 (300 topics—Chinese version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | 阿里, 日起, 黄河, 优惠, 公园, 顶级, 一幅, 大自然, 夜晚, 新篇 | 自然景观 |
| Topic 2 | 时代, 记录, 伴随, 短暂, 流动, 一代, 太快, 快到, 奔涌, 变迁 | 变迁 |
| Topic 3 | 朋友, 蔬菜, 一份, 大盘, 拒绝, 满满, 颜值, 蓝莓, 洋葱, 见面 | 饮食 |
| Topic 4 | 家族, 创作, 科学, 遇见, 分子, 符号, 作者, 融入, 象征, 数学 | 创作 |
| Topic 5 | 清单, 不行, 心碎, 一把, 本届, 相见, 闭幕式, 分手, 组合体, 欧豪 | 活动 |
| Topic 6 | 地点, 音乐, 特色, 首个, 错过, 宝藏, 音乐节, 景区, 人文, 闪亮 | 音乐节 |
| Topic 7 | 居民, 中午, 在线, 阵阵, 公安, 厨房, 赶到, 报警, 众人, 热心 | 社区服务 |
| Topic 8 | 调查, 一段, 睡觉, 关键, 歌曲, 大脑, 反复, 资源, 效应, 旋律 | 大脑研究 |
| Topic 9 | 学习, 评论, 近日, 条件, 意见, 网传, 公众, 仪式, 简称, 发布公告 | 舆论 |
| Topic 10 | 工作, 第一次, 办公室, 办公, 一周, 工作人员, 馋嘴, 苦涩, 小小, 深刻 | 工作 |

**Table 5:** Sample topics generated by LDA 150 (150 topics—English version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | Invitation, enhanced, hosting, competition, under, moment, overall, master, Chinese-style, beauty | Beauty event |
| Topic 2 | Future, donation, industry, team, agriculture, solid-state, battery, counterattack, project, cultivate | Industrial development |
| Topic 3 | Content, Saturday, only, reading, keep away from, topic, temperament, heard, Northeast, program | – |

(Continued)

**Table 5 (continued)**

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 4 | Still, citizens, pause, years, cabinet, public, that day, completed, presented, square | Urban construction |
| Topic 5 | Research, advanced, complete, director, negotiation, applause, life, watched, myth, two films | Movie |
| Topic 6 | Protection, drone, attack, many places, driver, government, privileges, a project, a group, transaction | Safety and policy |
| Topic 7 | Temporarily, wartime, manual, care, open, century, rest, a root, surprised, interview | Wartime response |
| Topic 8 | Tell, estimate, risk, number, enhance, deny, group, create, together, consultation | Risk management |
| Topic 9 | Peru, surgery, expert, brand, smile, large-scale, two people, suspect, salute, one hand | Medical |
| Topic 10 | City, office, guarding, winter, Fujian, three years, observation, rhythm, frontline, beauty | City |

**Table 6:** Sample topics generated by LDA 300 (300 topics—English version)

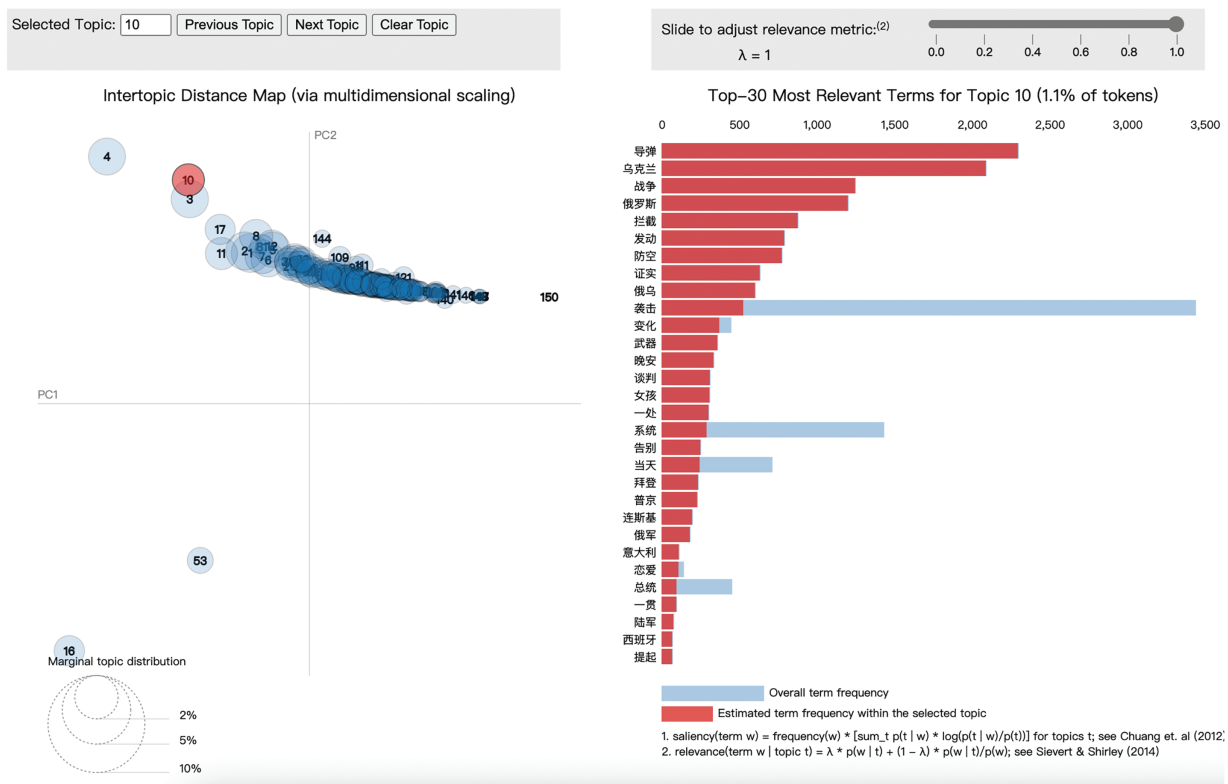| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | Alibaba, sunrise, Yellow River, discount, park, top-level, a picture, nature, night, new chapter | Natural landscape |
| Topic 2 | Era, record, accompany, brief, flow, one generation, too fast, surging, changes | Change |
| Topic 3 | Friends, vegetables, a dish, large plate, refuse, full, beauty, blueberry, onion, meet | Diet |
| Topic 4 | Family, creation, science, encounter, molecule, symbol, author, integrate, symbol, mathematics | Creation |
| Topic 5 | List, not allowed, heartbreak, a bunch, this edition, meet, closing ceremony, breakup, combination, Hao OU | Event |
| Topic 6 | Location, music, feature, first, missed, treasure, music festival, scenic spot, culture, shining | Music festival |
| Topic 7 | Residents, noon, online, waves, police, kitchen, rush to, report, crowd, warm-hearted | Community service |
| Topic 8 | Survey, a section, sleep, key, song, brain, repeat, resource, effect, melody | Brain research |
| Topic 9 | Study, comments, recently, condition, opinion, online, public, ceremony, abbreviation, announcement | Public opinion |
| Topic 10 | Work, first time, office, work, one week, staff, gluttonous, bitter, small, profound | Work |

**Figure 4:** Visual representation of LDA-generated topics (150 topics)

### 4.3.2 BERTopic Results

In contrast to LDA, BERTopic requires much less data preprocessing, does not require complex operations (e.g., elimination of stop words) as in LDA, and does not require the number of topics to be specified in advance. We have trained the BERTopic model using the default settings. Like LDA, social researchers have examined, interpreted, and annotated the generated topic samples. Table 7 shows a sample of 10 topics, and Table 8 presents the corresponding English translation. Figs. 5 and 6 illustrate the keyword c-TF-IDF scores for a sample of topics in Chinese and English, respectively. These scores reflect the importance of each topic word in the original texts. BERTopic outperforms LDA in terms of the interpretability of the generated topics. A plot of inter-topic distances is shown in Fig. 7, demonstrating the statistical proximity of the individual topics.

**Table 7:** Topics generated by BERTopic (Chinese version)

| No. | Top-10 Keywords | Annotation |
|---------|-----------------------------------------------------------|:----------:|
| Topic 1 | 美食, 好吃, 好物, 闭眼, 口感, 火锅, 美味, 一口, 超级, 苹果 | 美食 |
| Topic 2 | cn, http, 麻子, 整人, 头条, kidding, 吃秀, 网站, 思考, 吃播 | – |
| Topic 3 | 微笑, 太好了, 喜欢, 对不起, 苦涩, 没干, 下跪, 救救, 不想, 许凯 | – |
| Topic 4 | 王源, 演唱会, 音乐, 不息, 声生, 新歌, 舞台, 肖战, 狂欢, 檀健次 | 明星 |
| Topic 5 | 以色列, 冲突, 加沙, 巴以, 黎巴嫩, 黎以, 真主党, 巴勒斯坦, 袭击, 以军 | 军事 |
| Topic 6 | 老师, 同学, 学习, 小蜜蜂, 考研, 学生, 孩子, 英语, 视频, 校园 | 教育 |
| Topic 7 | cn, http, 成毅, 肖战, cy, 超话, 云弘深, 00, 少年, 樊振东 | – |

(Continued)

**Table 7 (continued)**

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 8 | 指数, 市场, 行情, 个股, 大盘, 股市, 反弹, 下跌, 资金, 上涨 | 经济 |
| Topic 9 | 乌克兰, 俄罗斯, 俄乌, 导弹, 弹道导弹, 洲际, 乌军, 发射, 局势, 俄军 | 军事 |
| Topic 10 | 减肥, 体重, 减重, 脂肪, 饮食, 减脂, 肥胖, 鲁肽, 美格, 运动 | 健康 |

**Table 8:** Topics generated by BERTopic (English version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | Food, Delicious, Good Stuff, Blindly Recommend, Texture, Hot Pot, Tasty, A Bite, Super, Apple | Food |
| Topic 2 | cn, http, Mazi, Prank, Headlines, Kidding, Eating Show, Website, Thoughts, Eating Show | – |
| Topic 3 | Smile, Great, Like, Sorry, Bitterness, Didn't Complete, The Kneeling, Help, Don't Want, Xu Kai | – |
| Topic 4 | Roy Wang, Concert, Music, Unceasing, Sound Life, New Song, Stage, Xiao Zhan, Carnival, Tan Jianci | Celebrity |
| Topic 5 | Israel, Conflict, Gaza, Israel-Palestinian, Lebanon, Lebanon-Israel, Hezbollah, Palestine, Attack, Israeli Army | Military |
| Topic 6 | Teacher, Classmate, Study, Little Bee, Postgraduate Exam, Student, Child, English, Video, Campus | Education |
| Topic 7 | cn, http, Cheng Yi, Xiao Zhan, cy, Super Topic, Yun Hongshen, 00, Youth, Fan Zhendong | – |
| Topic 8 | Index, Market, Trends, Stocks, Main Board, Stock Market, Rebound, Decline, Funds, Rise | Economics |
| Topic 9 | Ukraine, Russia, Russia-Ukraine, Missile, Ballistic Missile, Intercontinental, Ukrainian Army, Launch, Situation, Russian Army | Military |
| Topic 10 | Weight Loss, Body Weight, Losing Weight, Fat, Diet, Fat Reduction, Obesity, Semaglutide, Wegovy, Exercise | Health |



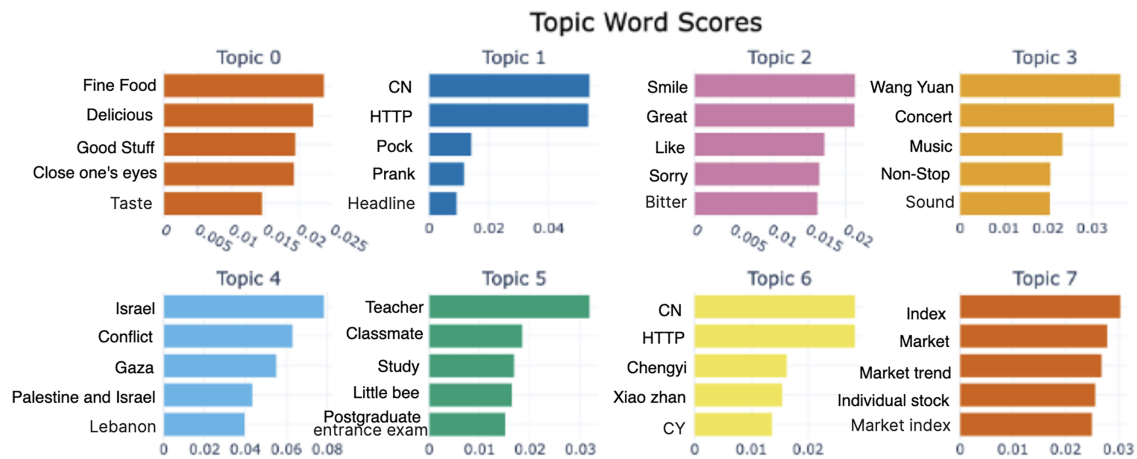**Figure 5:** Topic words by the BERTopic model

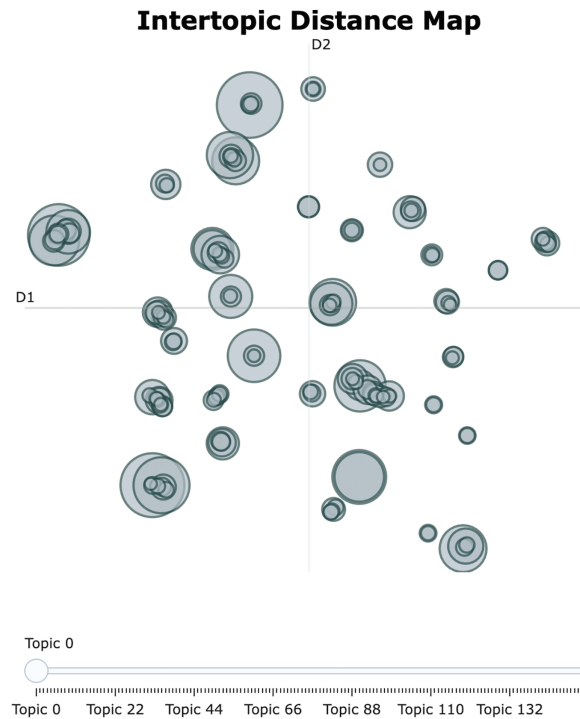**Figure 6:** Topic words by the BERTopic model (Enlgish translation)



**Figure 7:** Visual presentation of BERTopic-generated topics

Unlike LDA, BERTopic does not require the number of topics to be generated to be specified in advance. What is unique is that BERTopic provides a hierarchical tree structure graph based on topic similarity, which helps social science researchers more clearly observe topic categories, as shown for the Chinese and English translate in Figs. 8 and 9, respectively.
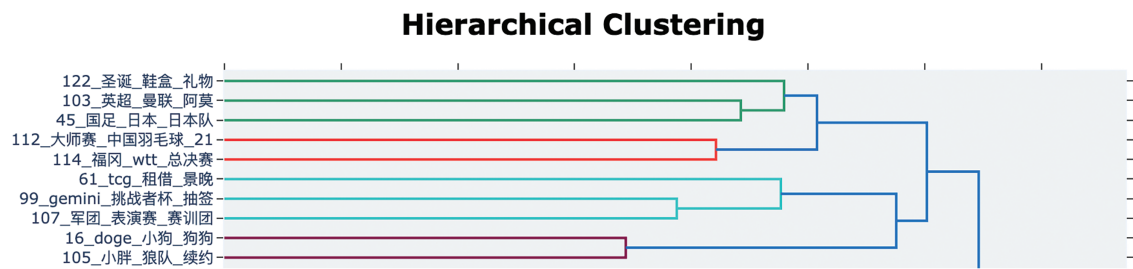
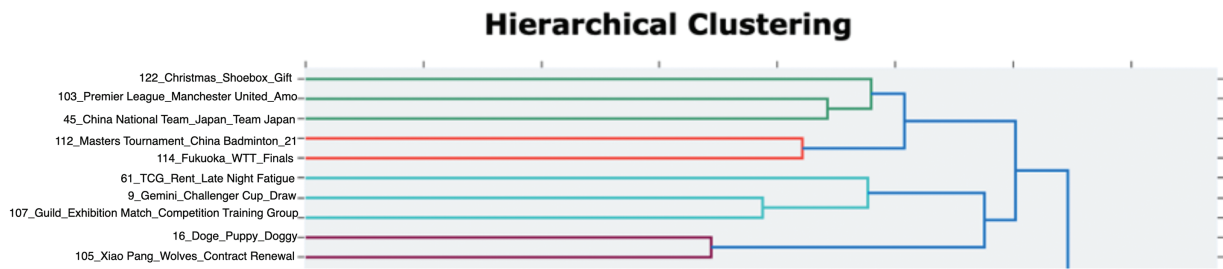**Figure 8:** Part of hierarchical topic clustering by the BERTopic



**Figure 9:** Part of hierarchical topic clustering by the BERTopic (English translation)

### 4.3.3 GLMTopic Results

Akin to BERTopic, the GLMTopic model can determine topics automatically without manually specifying the number of topics. Its innovation is that GLMTopic extracts topic words more relevant to the text content based on the semantic association between the weighted word frequency matrix and the original text, adopting the GLM4-Long model (see Table 9 for the Chinese version and Table 10 for the English translation). This approach makes the topics more contextualized, and compared to LDA and BERTopic, GLMTopic has a significant advantage in the interpretability and accuracy of the generated topics.

**Table 9:** Topics generated by GLMTopic (Chinese version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | 股市, 机器人, 无人机, 人工智能, 固态电池, 新能源, G, 云计算, 量子计算, 区块链 | 科技 |
| Topic 2 | 互联网生活, 网络社交, 在线娱乐, 网络文化, 网络消费, 网络言论, 网络隐私, 网络安全, 网络道德, 网络监管 | 互联网 |
| Topic 3 | 猫咪, 萌宠, 宠物, 养猫, 养狗, 宠物护理, 宠物健康, 宠物食品, 宠物玩具, 宠物训练 | 宠物 |
| Topic 4 | 俄乌冲突, 乌克兰导弹袭击, 俄罗斯核学说, 西方武器支持, 乌克兰难民危机, 美国对乌政策, 乌克兰经济发展, 乌克兰文化传承, 乌克兰社会问题, 乌克兰未来展望 | 军事 |
| Topic 5 | 生育率, 韩国生育, 生育政策, 再婚生育, 双胞胎性格, 情绪影响, 健康小知识, 女性健康, 母婴育儿, 孕产知识 | 社会 |
| Topic 6 | 中国航展, 珠海航展, 歼-A, 歼-T, 空军八一飞行表演队, 苏-, 谢尔盖, 虞书欣, 丁禹兮, 飞行表演 | 军事 |

(Continued)

**Table 9 (continued)**

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 7 | 微博, 章昊, 永夜星河, 哈哈哈, 分享一个抽象可以整人的小网站和制作, ankvn, andwwmw, 非主流闪光字体之类的, 防地震建筑设计比赛 | 娱乐 |
| Topic 8 | 婚礼摄影, 婚礼策划, 婚礼现场, 婚礼跟拍, 婚礼预告, 婚礼习俗, 婚礼布置, 婚礼化妆, 婚礼服装, 婚礼音乐 | 婚礼 |
| Topic 9 | 房地产市场, 房产税减免, 住房交易税费, 契税增值税, 普通住房标准, 非普通住房标准, 楼市政策, 房地产新政, 楼市调控, 房价走势 | 房地产 |
| Topic 10 | 穷富差距, 社会阶层, 人际关系, 社会不公, 情感纠葛, 生活困境, 心理健康, 人性弱点, 生活态度, 个人成长 | 社会问题 |

**Table 10:** Topics generated by GLMTopic (English version)

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 1 | Stock Market, Robotics, Drones, Artificial Intelligence, Solid State Batteries, New Energy, G, Cloud Computing, Quantum Computing, Blockchain | Technology |
| Topic 2 | Internet Life, Online Social Networking, Online Entertainment, Online Culture, Online Consumption, Online Speech, Online Privacy, Online Safety, Online Ethics | Internet |
| Topic 3 | Cats, Cuddly Pets, Pets, Cats, Dogs, Pet Care, Pet Health, Pet Food, Pet Toys, Pet Training | Pets |
| Topic 4 | Russia-Ukraine conflict, Ukrainian missile attacks, Russian nuclear doctrine, Western weapons support, Ukrainian refugee crisis, U.S. policy toward Ukraine, Ukrainian economic development, Ukrainian cultural heritage, Ukrainian social issues, Ukraine's outlook | Military |
| Topic 5 | Birth Rate, South Korea Birth Rate, Birth Policy, Remarriage and Childbirth, Twin Personality, Emotional Impact, Health Tips, Women's Health, Maternal and Infant Care, Pregnancy Knowledge | Society |
| Topic 6 | Airshow China, Zhuhai Airshow, J-A, J-T, Air Force August 1st Air Show Team, Su-, Sergey, Yu Shuxin, Ding Yuxi, Air Show | Military |
| Topic 7 | Weibo, Zhang Hao, Eternal Night Galaxy, Hahaha, Sharing an Abstract Website for Fun, ankvn, andwwmw, Non-mainstream Flash Fonts, Earthquake-resistant Building Design Competition | Entertainment |
| Topic 8 | Wedding Photography, Wedding Planning, Wedding Venues, Wedding Videography, Wedding Preview, Wedding Traditions, Wedding Decorations, Wedding Makeup, Wedding Attire, Wedding Music | Wedding |
| Topic 9 | Real Estate Market, Property Tax Reduction, Housing Transaction Fees, Deed Tax and Value-Added Tax, Standard Housing Criteria, Non-Standard Housing Criteria, Real Estate Policies, New Real Estate Policies, Housing Market Regulation, Housing Price Trends | Real Estate |

**Table 10 (continued)**

| No. | Top-10 Keywords | Annotation |
|---|---|---|
| Topic 10 | Wealth Gap, Social Classes, Interpersonal Relationships, Social Inequality, Emotional Entanglements, Life Struggles, Mental Health, Human Weaknesses, Life Attitudes, Personal Growth | Social Issues |

GLMTopic provides distance plots between different topics to show the statistical similarity of topics and supports the output of word frequency matrix scores and topic hierarchical plots, as shown in Figs. 10–12 (English translate), which demonstrate the first 40 topics because of the large number of topics in this paper. GLMTopic also generates topic word clouds (see Fig. 13), facilitating social researchers' understanding of the topics more intuitively.
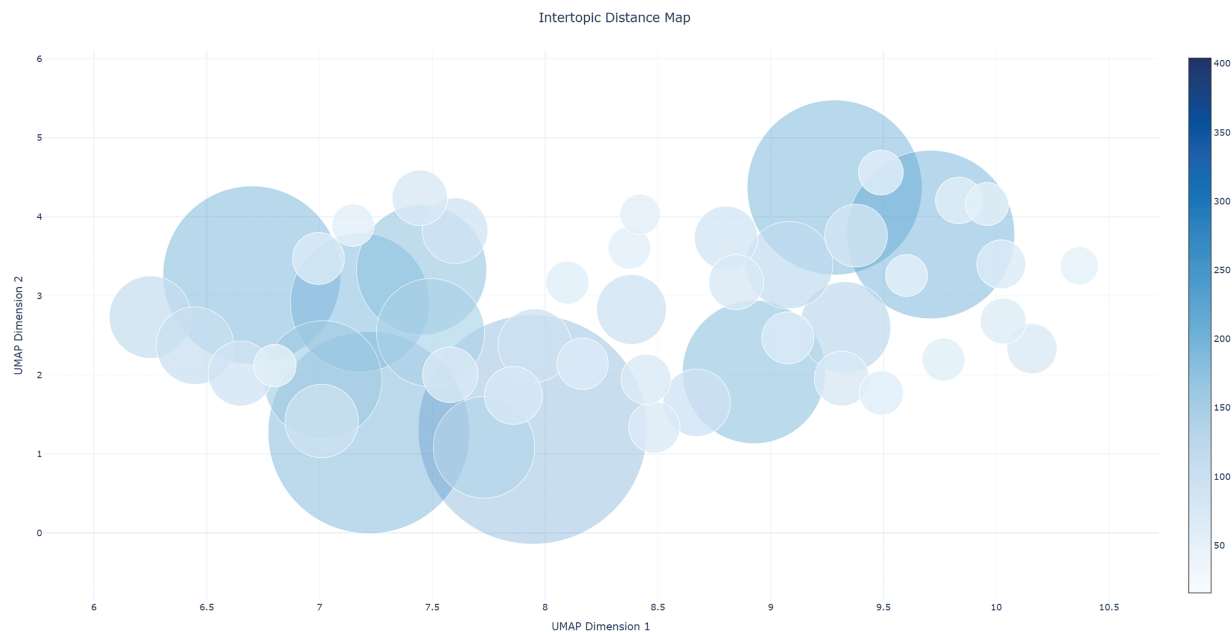


**Figure 10:** Visual display of topics generated by GLMTopic (top 50 topics)
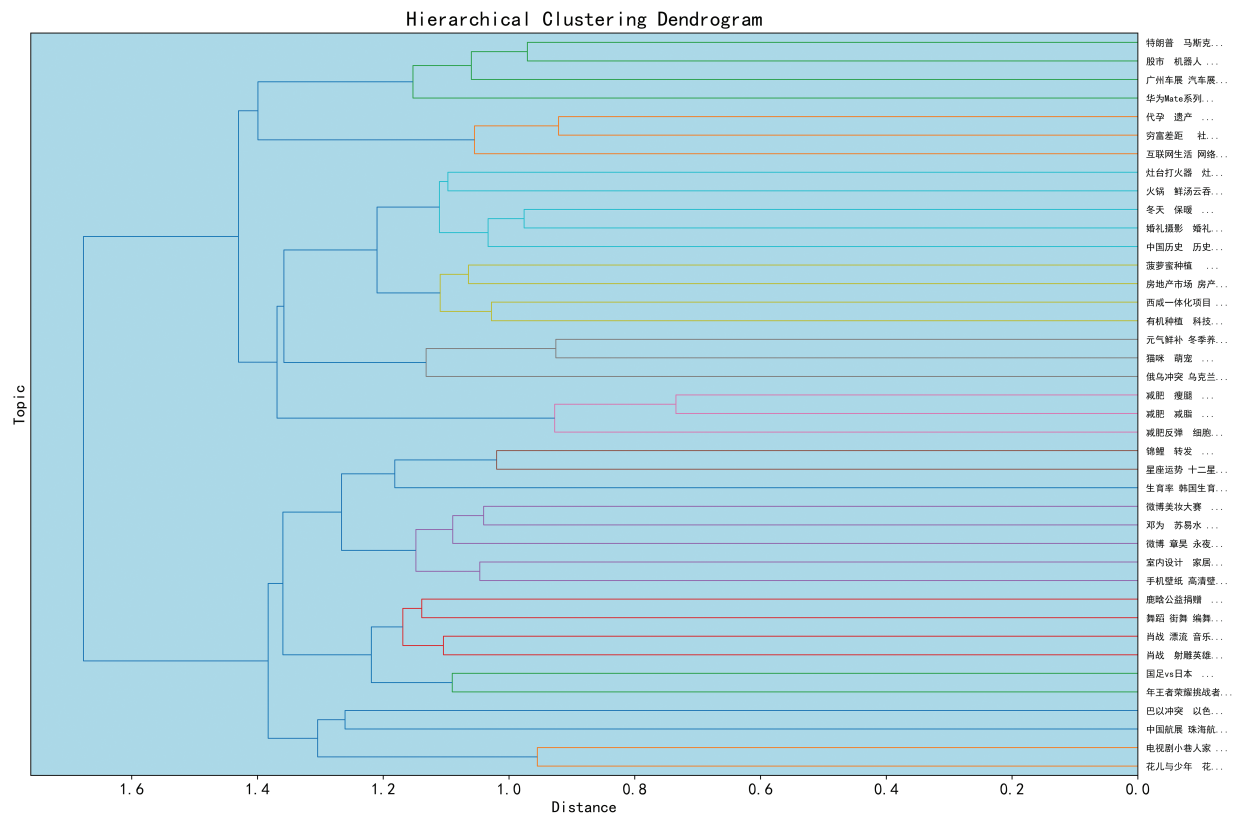
**Figure 11:** Tree diagram of relationships between topics generated by GLMTopic (top 40 topics)
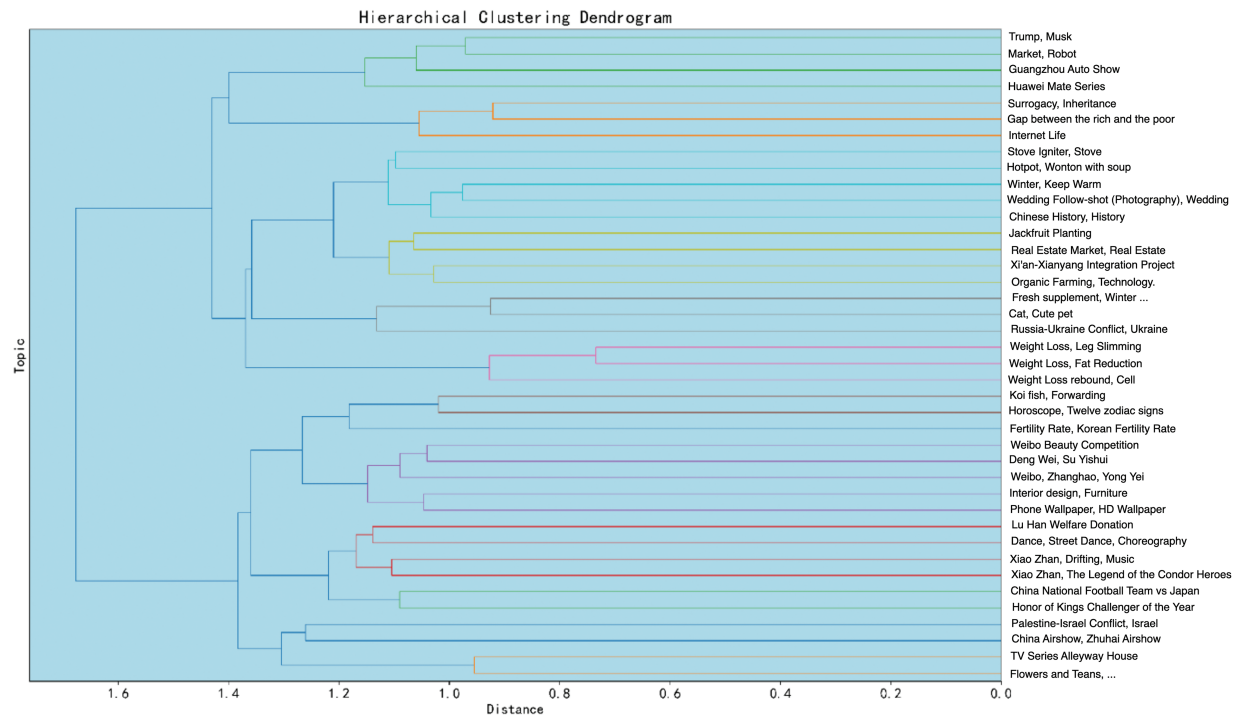


**Figure 12:** Tree diagram of relationships between topics generated by GLMTopic (top 40 topics, English translate)

**Figure 13:** GLMTopic generates a word cloud of Topic 1 (technology)

## 4.4 Evaluation of Topic Models

When evaluating topic models, the topic coherence score is considered one of the most appropriate evaluation metrics when the output of the topic model is intended for use by human users. The topic coherence score reflects the quality of a topic by measuring the consistency between keywords in the topic. It is a widely used and essential evaluation method in topic modeling. Among the calculation methods of topic coherence scores, c_v is the most used metric, which constructs content vectors based on the co-occurrence relationships of keywords and calculates the score by combining cosine similarity and normalized point mutual information (NPMI) to measure the internal consistency of the keywords in a topic. Although topic coherence scores are critical evaluation criteria, they are not perfect. In this study, to comprehensively evaluate the effectiveness and efficiency of topic modeling, we combined the following three methods:

(1) Coherence (c_v score): to quantitatively evaluate the quality of topics generated by the model. In this experiment, topic coherence scores were computed using the Gensim Python library.
(2) Interpretability: to involve human experts to verify the actual interpretability and reliability of the generated topics. We recruited three independent social science researchers and performed an inter-expert reliability analysis to confirm the results.
(3) Level of human involvement: to involve human experts to assess the level of human involvement in the manual label annotation of generated topics.

### 4.4.1 Coherence

Fig. 14 compares the c_v scores of the topics generated by each model. The experimental results show that GLMTopic, which incorporates the GLM4-Long model, is superior to other models in topic coherence scores in the Chinese short text topic modeling scenario. The strong contextual reasoning ability of the GLM4-Long model allows GLMTopic to more accurately capture the semantics of short texts in the Chinese context, thus generating more explanatory and coherent topics. Regarding c_v scores, GLMTopic scores are 0.610, significantly higher than LDA300 (up to 0.598) and BERTopic (0.585), highlighting the higher consistency of topics with LLM tuning.
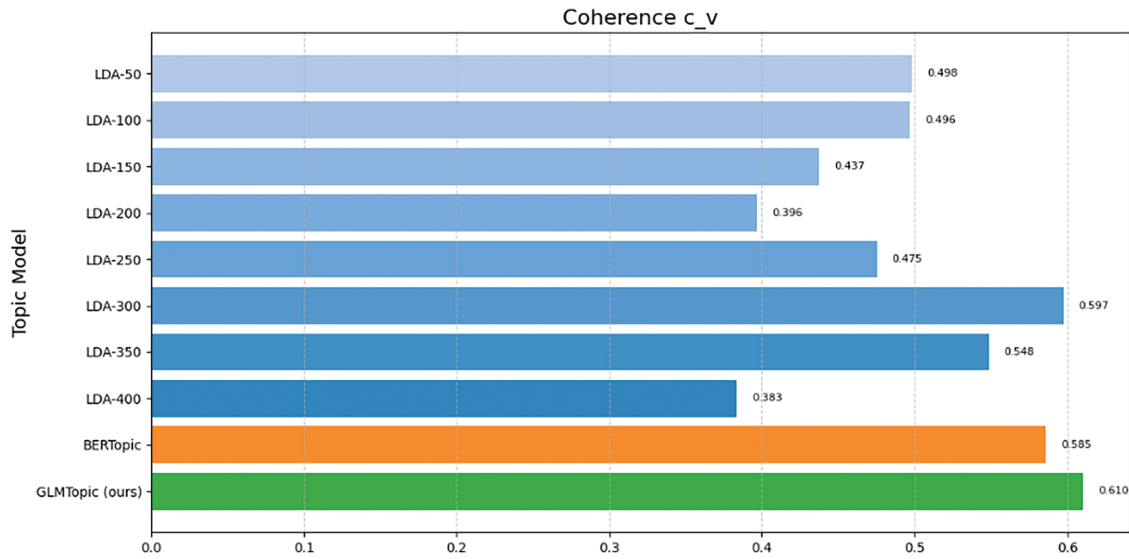
**Figure 14:** Comparison of various topic models used in the experiment by the c_v coherence score

To verify statistical significance on the c_v coherence score, we adopted the bootstrap method combined with a paired $t$-test for statistical verification. The method is as follows: We construct a simulated distribution for each model by taking 50 samples from the observed c_v scores of each model and assuming that they fit a normal distribution with a mean of the score and a standard deviation of 0.02. We then tested the mean difference between the GLMTopic model and the baseline model (BERTopic and LDA-300) using a $t$-test. The results show that:

(1)    GLMTopic vs. BERTopic: $t = 194.19$, $p < 0.001$;
(2)    GLMTopic vs. LDA-300: $t = 101.01$, $p < 0.001$.

This indicates that GLMTopic performs better with significance than the baseline model in terms of topic consistency. Fig. 15 shows the distribution of topic Coherence (c_v) scores for GLMTopic, BERTopic, and LDA-300 over 1000 Bootstrap resampling.

*4.4.2 Interpretability and Level of Human Involvement*

For the evaluation of interpretability, the three social science experts we recruited were involved in the annotation and assessment process. To assess the consistency of the interpretability evaluation, we conducted an inter-rater reliability analysis among the three independent social science experts. The three social science experts comparatively evaluated the interpretability of the GLMTopic, BERTopic, and LDA models by explicitly ranking them from most interpretable (Rank 1) to least interpretable (Rank 3), as shown in Table 11. Experts unanimously agreed that GLMTopic offered superior interpretability. The calculated Fleiss' Kappa value ($\kappa = 0.33$) indicates fair overall agreement, reflecting slight disagreements in interpretability rankings between BERTopic and LDA.

Each expert independently labeled the topics generated by LDA and BERTopic based on their relevance and coherence. All three social science researchers have confirmed that GLMTopic improved intuitiveness and minimizes manual interpretation, and no manual annotation was required after the topic generation. Table 12 extends the performance comparison, illustrating the coherence and level of human involvement.
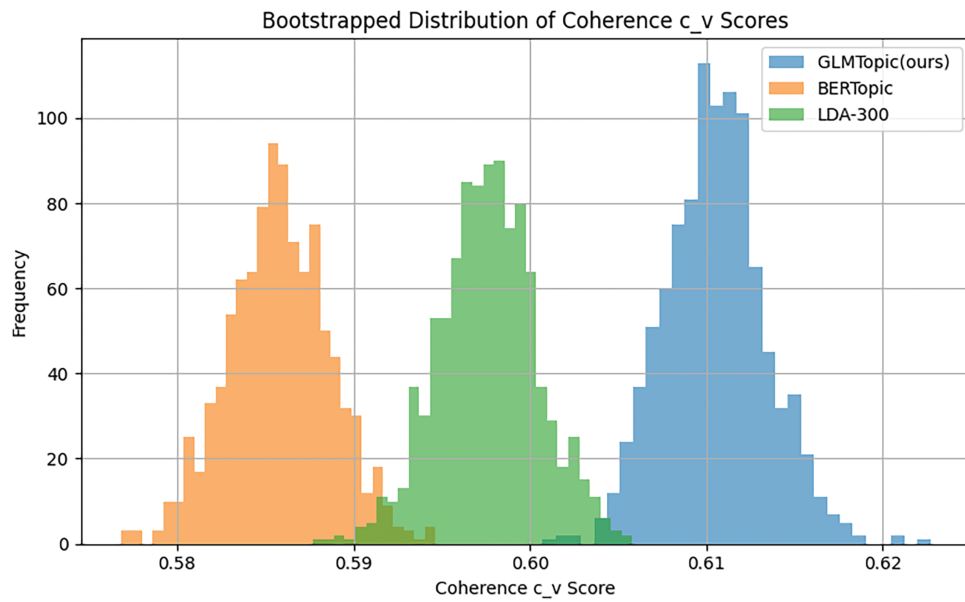
**Figure 15:** Distribution of coherence scores for topic models over 1000 Bootstrap resamples

**Table 11:** Expert rankings of model interpretability

| Expert | LDA | BERTopic | GLMTopic |
|--------|-----|----------|----------|
| 1 | 3 | 2 | 1 |
| 2 | 2 | 3 | 1 |
| 3 | 3 | 2 | 1 |

**Table 12:** Comparison between selected topic modeling models

| Topic model. | No. of topic | Coherence (c_v) | Level of human involvement |
|--------------|--------------|-----------------|----------------------------|
| LDA50 | 50 | 0.498 | |
| LDA100 | 100 | 0.496 | |
| LDA150 | 150 | 0.437 | |
| LDA200 | 200 | 0.396 | High |
| LDA250 | 250 | 0.475 | |
| LDA300 | 300 | 0.597 | |
| LDA350 | 350 | 0.548 | |
| LDA400 | 400 | 0.383 | |
| BERTopic | 151 | 0.585 | Low |
| GLMTopic (ours) | 385 | 0.610 | None |

## *4.5 Key Findings*

The key findings of this experiment are as follows.

**Data preparation:** LDA models require more complex data preprocessing, including non-trivial steps such as punctuation and stop-word removal, text normalization, text segmentation, etc., often requiring considerable time and resource consumption. By contrast, BERTopic and GLMTopic are equally good at data preparation and require minimal preprocessing. The underlying neural networks enable them to capture contextual relationships.

**Coherence:** In the experiment, GLMTopic performed best, with a significantly higher coherence score than the other state-of-the-art models. This indicates that GLMTopic can generate more semantically consistent topics that are more in line with the actual content of the text than the different models. By integrating ACGE and GLM, GLMTopic can effectively utilize the contextual information in short texts to model the text semantics in-depth, thus generating more in-depth and meaningful topics. ACGE extracts high-quality semantic embeddings from the original text by combining the advantages of community structure and graph neural networks. GLMTopic breaks through the limitations of the traditional embedding model by enhancing the embedding through ACGE's community, which optimizes the semantic representation and makes the topic modeling results more closely fit the actual text content. It optimizes the semantic representation and makes the topic modeling results more closely match the actual text content.

In the context of Chinese short text scenarios, BERTopic is inferior to GLMTopic because the quality of the embedding model is limited by the corpus used for training, so BERTopic does not work as well for Chinese as GLMTopic. While not as good as GLMTopic, BERTopic performs better than most LDA configurations. Still, its performance is limited by the quality of the embedding model and short text adaptation, which does not fully utilize the contextual information. On the other hand, the coherence score of LDA is relatively low and parameter-sensitive, requiring researchers to constantly tune the parameters.

**Interpretability and level of human involvement:** GLMTopic excels in data preparation, topic coherence, and interpretability through its fusion of ACGE (Adaptive Community-enhanced Graph Embedding) and GLM. In the data preparation stage, GLMTopic utilizes ACGE's community-enhanced embedding technology to extract high-quality semantic embeddings directly from the original text without the need for complex data cleaning and word segmentation operations, which significantly improves processing efficiency and adaptability and is more advantageous than traditional LDA and BERTopic. Regarding topic generation, GLMTopic not only achieves a higher topic coherence score but also digs deeper into the semantic nature of the text through GLM's powerful contextual understanding capability so that the generated topics are more relevant to the actual content and have a higher degree of interpretability. In contrast, although BERTopic performs well in some scenarios, due to its dependence on external embedding models (e.g., Sentence-BERT), it is not effective enough in dealing with short Chinese texts due to the quality of embedding models and context modeling capability. The LDA model not only requires tedious data preprocessing and parameter debugging but also generates topics that are difficult to compare with GLMTopic and BERTopic regarding topic coherence and semantic consistency. Overall, GLMTopic performs optimally in the topic modeling task of Chinese short text and demonstrates strong semantic understanding and modeling ability through the organic combination of ACGE and GLM, providing an efficient and reliable solution for complex text analysis.

## 5 Conclusions and Future Work

In this paper, we have introduced GLMTopic, a novel hybrid topic modeling framework that leverages large language models to enhance topic discovery, coherence, and interpretability, particularly in the context of Chinese short-text datasets. We evaluated its performance against two established topic modeling approaches, namely LDA and BERTopic, and demonstrated that GLMTopic outperforms both in terms of topic coherence and usability while also reducing the need for extensive text preprocessing.

This study has successfully addressed the two research questions outlined in Section 1.2:

**RQ 1: Can we leverage the capabilities of LLMs to improve Chinese topic modeling performance for social science research on large-scale social media data?**

The experimental results confirm that our proposed hybrid model, GLMTopic, integrates LLMs into topic modeling and can result in more coherent and contextually relevant topics in the Chinese context. By combining ACGE, dimensionality reduction with UMAP, and clustering via HDBSCAN, representation tuning with GLM, GLMTopic has achieved superior performance over state-of-the-art methods in terms of coherence. This hybrid approach demonstrates the potential of LLM-enhanced topic modeling as a more effective and scalable solution.

**RQ 2: How can we interpret the topic modeling results without intervention from human experts to enhance their applicability?**

GLMTopic significantly improves the interpretability of topic modeling results by automating the generation of meaningful and contextually appropriate topic labels. By integrating GLM for representation tuning, the framework refines extracted topics to align with the original text, reducing ambiguity and improving readability. The results indicate that this approach minimizes human intervention, making topic modeling more accessible for researchers, analysts, and practitioners. Furthermore, interactive visualizations enhance the usability of the generated topics, providing an intuitive way to explore topic structures without requiring domain expertise.

Despite its advantages, this study has certain limitations. The reliance on LLMs introduces computational costs, which may limit scalability in real-time or resource-constrained environments. Additionally, while topic coherence is a widely used evaluation metric, alternative metrics such as thematic diversity could be explored to refine topic quality assessments further. Future research will focus on optimizing computational efficiency by experimenting with more lightweight and efficient LLM architectures and expanding the model's applicability to real-time analysis and cross-lingual topic modeling. Moreover, integrating knowledge graphs and sentiment analysis could further enhance the contextual relevance of extracted topics, making GLMTopic a more comprehensive tool for social media analysis, news insights, and other high-impact applications.

Overall, this study highlights the transformative potential of LLM-powered topic modeling in improving topic discovery and interpretation, paving the way for more efficient, automated, and semantically rich text analysis pipelines.

**Author Contributions:** The authors confirm their contribution to the paper as follows: conceptualization, Weisi Chen; methodology, Weisi Chen and Walayat Hussain; software, Weisi Chen and Junjie Chen; validation, Weisi Chen and Walayat Hussain; formal analysis, Junjie Chen and Weisi Chen; investigation, Walayat Hussain; resources, Weisi Chen; data curation, Weisi Chen and Junjie Chen; writing—original draft preparation, Weisi Chen and Junjie Chen; writing—review and editing, Weisi Chen and Walayat Hussain; visualization, Walayat Hussain and Junjie Chen; supervision, Weisi Chen; project administration, Weisi Chen and Walayat Hussain; funding acquisition, Weisi Chen and Walayat Hussain. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data supporting this study's findings are available from the Corresponding Authors, Weisi Chen and Walayat Hussain, upon reasonable request.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. Inf Syst. 2023;112:102131. doi:10.1016/j.is.2022.102131.
2. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
3. Lamirel J-C, Lareau F, Malaterre C. CFMf topic-model: comparison with LDA and Top2Vec. Scientometrics. 2024;129(10):6387–405. doi:10.1007/s11192-024-05017-z.
4. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. 2022.
5. Madzik P, Falat L, Jum'a L, Vrábliková M, Zimon D. Human-centricity in Industry 5.0—revealing of hidden research topics by unsupervised topic modeling using Latent Dirichlet Allocation. Eur J Innov Manag. 2025;28(1):113–38. doi:10.1108/ejim-09-2023-0753.
6. Ahammad T. Identifying hidden patterns of fake COVID-19 news: an in-depth sentiment analysis and topic modeling approach. Nat Lang Process J. 2024;6(6):100053. doi:10.1016/j.nlp.2024.100053.
7. Rabadán-Martín I, Barcos-Redín L, Pereira-Delgado J, Aguado-Correa F, Padilla-Garrido N. Topic-based engagement analysis: focusing on hotel industry Twitter accounts. Tour Manag. 2025;106(4):104981. doi:10.1016/j.tourman.2024.104981.
8. Bhat MR, Kundroo MA, Tarray TA, Agarwal B. Deep LDA: a new way to topic model. J Inf Optim Sci. 2020;41(3):823–34. doi:10.1080/02522667.2019.1616911.
9. Chen W, Rabhi F, Liao W, Al-Qudah I. Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: a comparative study. Electronics. 2023;12(12):2605. doi:10.3390/electronics12122605.
10. Pratiwi MD, Tania KD. Knowledge discovery through topic modeling on GoPartner user reviews using BERTopic, LDA, and NMF. J Appl Inform Comput. 2025;9(1):1–7. doi:10.30871/jaic.v9i1.8782.
11. Chen J, Chen W, Zheng J, Xie Q. Employing topic and sentiment language models for news analysis on the COVID-19 Pandemic. In: Proceedings of the 2024 International Conference on Intelligent Education and Computer Technology; 2024 Jun 28–30; Guilin, China.
12. Tang Z, Pan X, Gu Z. Analyzing public demands on China's online government inquiry platform: a BERTopic-Based topic modeling study. PLoS One. 2024;19(2):e0296855. doi:10.1371/journal.pone.0296855.
13. Ramamoorthy T, Kulothungan V, Mappillairaju B. Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. Front Artif Intell. 2024;7:1329185. doi:10.3389/frai.2024.1329185.
14. Wijanto MC, Widiastuti I, Yong H-S. Topic modeling for scientific articles: exploring optimal hyperparameter tuning in BERT. Int J Adv Sci Eng Inf Technol. 2024;14(3):912–9. doi:10.18517/ijaseit.14.3.19347.
15. Li T, Cui L, Wu Y, Pandey R, Liu H, Dong J, et al. Unveiling and advancing grassland degradation research using a BERTopic modelling approach. J Integr Agric. 2025;24(3):949–65. doi:10.1016/j.jia.2024.11.008.
16. Lee Y-G, Kim S. A comparative study on topic modeling of LDA, Top2Vec, and BERTopic models using LIS journals in WoS. J Korean Soc Libr Inf Sci. 2024;58(1):5–30.
17. Raman R, Pattnaik D, Hughes L, Nedungadi P. Unveiling the dynamics of AI applications: a review of reviews using scientometrics and BERTopic modeling. J Innov Knowl. 2024;9(3):100517. doi:10.1016/j.jik.2024.100517.
18. Zhu E, Yen J. BERTopic-driven stock market predictions: unraveling sentiment insights. arXiv:2404.02053. 2024.
19. Chen W, Hussain W, Al-Qudah I, Al-Naymat G, Zhang X. Intersection of machine learning and mobile crowdsourcing: a systematic topic-driven review. Pers Ubiquitous Comput. 2025;29(1):77–101. doi:10.1007/s00779-024-01820-w.
20. Zengul F, Bulut A, Oner N, Ahmed A, Yadav M, Gray HG, et al. A practical and empirical comparison of three topic modeling methods using a COVID-19 corpus: lSA, LDA, and Top2Vec. In: Proceedings of the 2023 Hawaii International Conference on System Sciences; 2023 Jan 3–6; Maui, HI, USA.
21. Yu D, Xiang B. Discovering topics and trends in the field of Artificial Intelligence: using LDA topic modeling. Expert Syst Appl. 2023;225(1):120114. doi:10.1016/j.eswa.2023.120114.

22. Yu D, Fang A, Xu Z. Topic research in fuzzy domain: based on LDA topic modelling. Inf Sci. 2023;648(12):119600. doi:10.1016/j.ins.2023.119600.

23. Uthirapathy SE, Sandanam D. Topic modelling and opinion analysis on climate change twitter data using LDA and BERT model. Procedia Comput Sci. 2023;218(1):908–17. doi:10.1016/j.procs.2023.01.071.

24. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J. 2024;6(6):100048. doi:10.1016/j.nlp.2023.100048.

25. GLM T, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. arXiv:2406.12793. 2024.

26. Wang H, Prakash N, Hoang NK, Hee MS, Naseem U, Lee RK-W. Prompting large language models for topic modeling. In: Proceedings of the 2023 IEEE International Conference on Big Data (BigData); 2023 Dec 15–18; Sorrento, Italy.

27. Chagnon E, Pandolfi R, Donatelli J, Ushizima D. Benchmarking topic models on scientific articles using BERTeley. Nat Lang Process J. 2024;6:100044. doi:10.1016/j.nlp.2023.100044.

28. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al., editors. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning; 2021 Jul 18–24; Online.

29. Li X, Li L, Jiang Y, Wang H, Qiao X, Feng T, et al. Vision-language models in medical image analysis: from simple fusion to general large models. Inf Fusion. 2025;118(2):102995. doi:10.1016/j.inffus.2025.102995.

30. Reimers N. Sentence-BERT: sentence embeddings using Siamese BERT-Networks. arXiv:1908.10084. 2019.

31. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. 2018.

32. Bhattacharjee P, Mitra P. A survey of density based clustering algorithms. Front Comput Sci. 2021;15(1):1–27. doi:10.1007/s11704-019-9059-3.