Advances in Computer and Communication, 2024, 5(1), 64-71
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Mongolian Topic Extraction Using Pre-trained Models

Ailiya, Qintu Si*, Siriguleng Wang

College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, Inner Mongolia, China.

How to cite this paper: Ailiya, Qintu Si, Siriguleng Wang. (2024) Mongolian Topic Extraction Using Pre-trained Models. *Advances in Computer and Communication*, 5(1), 64-71.
DOI: 10.26855/acc.2024.02.011

Received: February 16, 2024
Accepted: March 13, 2024
Published: April 10, 2024

***Corresponding author:** Qintu Si, College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, Inner Mongolia, China.

Abstract

The paper introduces a Mongolian topic extraction method that utilizes a pre-trained language model to improve the quality of Mongolian intelligent question answering. Initially, the Mongolian text data undergoes preprocessing steps, including text correction, data cleaning, and word segmentation, to ensure accurate and readable data. Stop words are then removed to reduce noise while filtering high- and low-frequency words to emphasize key terms for constructing a Mongolian thesaurus. After preprocessing, the pre-trained model is used to represent Mongolian word vectors that capture semantic meanings in the language. Based on these representations, an unsupervised topic extraction method employing a topic model identifies and clusters similar topics within the text, providing a structured representation of the data. Experimental results demonstrate that this proposed method outperforms traditional topic extraction methods such as latent Dirichlet allocation and embedded topic models with improvements of 0.3406 and 0.0675 in terms of topic extraction quality respectively, showcasing its effectiveness and efficiency in extracting relevant topics from Mongolian text. This approach enhances text comprehensibility.

Keywords

Mongolian, Topic model, Text semantic understanding, Pre-trained model

1. Introduction

The integration of the internet and big data technologies has significantly increased the importance of efficiently extracting key topics from extensive textual datasets, posing a formidable challenge in natural language processing. Topic modeling, an essential technique in NLP for uncovering hidden semantic structures, has been widely used. Approaches in this field can be categorized into traditional models, such as LDA [1] and PLSA [2] that cluster text semantics without supervision but may lack word correlations impacting precision and coherence, and deep learning models, like GLDA [3], GMNTM [4], neural network-augmented approaches [5] that combine word vector representations with neural architectures to capture nuanced semantic relationships. ETM [6] effectively handles large-scale text data by merging traditional topic modeling with word embeddings. BERTopic[7] by Grootendorst and Yuhan Wang et al. [8] is a versatile BERT-based embedded topic model developed for various datasets.

In the Mongolian context, Jiang Yupeng [9] introduced a Cyrillic Mongolian LDA topic model while Siriguleng [10] developed an LDA-based information retrieval system. Fu Yujiao [11] integrated topical information into a Variational Autoencoder (VAE) for Mongolian short texts, whereas Bai Shuguang [12] enhanced TextRank for keyword extraction. However, Mongolian topic modeling has predominantly relied on traditional approaches suffering from limited scalability and generality.

This paper presents an approach to extract topics in Mongolian using a pre-trained model's semantic understanding to identify relevant topics. The method involves text normalization, word segmentation, and stop word removal followed by generating semantic word vectors which are then utilized in an unsupervised clustering algorithm for topic modeling. Experimental evidence supports its superior performance.

2. Mongolian topic extraction method based on a pre-trained model

2.1 Data processing

To improve the training effect of the topic model, a series of data processing operations are needed before training the topic model. The specific steps are as follows:
1) Data cleaning, removing noise, and simplifying the text structure; obtaining 2.67 million training data.
2) Automatic text correction, using traditional Mongolian language proofreading software to enhance accuracy; increasing the proportion of correct words in each text group by 7.71%.
3) Word segmentation, adopting a partial cutting method to construct a suitable length Mongolian language vocabulary that avoids sparsity and unknown word issues.

2.2 Constructing a Mongolian Thesaurus

2.2.1 Construction of the stop-word list

The accuracy and efficiency of Mongolian topic extraction were improved by developing a customized stop-word list, effectively eliminating irrelevant frequently occurring words. In this study, the stopword list was constructed based on existing research [13] using word frequency statistics. The construction process involved:
1) Data Preparation and Preprocessing: We collected and cleansed a diverse Mongolian corpus from various domains to maintain data integrity.
2) Part-of-Speech Tagging: The BiLSTM-CRF part-of-speech tagging tool was applied to the Mongolian corpus, utilizing the initial level of Mongolian Word Tagging for Information Technology Information Processing, covering 19 part-of-speech categories.
3) Stop-word Filtering: Nouns and verbs were preserved, while other parts-of-speech were considered as stop words due to their lesser contribution to the topic.
4) Human Review and Adjustment: Automatically generated stop words were meticulously reviewed, evaluated, and refined by professionals to ensure precision in alignment with context and semantics.

The result is a comprehensive Mongolian stop-word list comprising 1,551 common entries crafted specifically for topic extraction.

2.2.2 Construction of a high-frequency word list

To construct a Mongolian thesaurus and improve topic extraction accuracy, steps were taken to process high-frequency words and exclude stop words:

Word Frequency Ranking: After removing stop words, topic words were sorted by frequency, highlighting their prevalence and significance in the text. This ranking enabled the identification of dominant terms and the creation of word clouds to visualize frequencies above 0.75%. As shown in Fig.1, these clouds exposed high-frequency words without thematic relevance, such as ᠬᠢᠭᠳᠡᠬᠦ (do), ᠪᠠᠢᠨ᠎ᠠ (exists), and ᠪᠣᠯᠤᠨ (become).

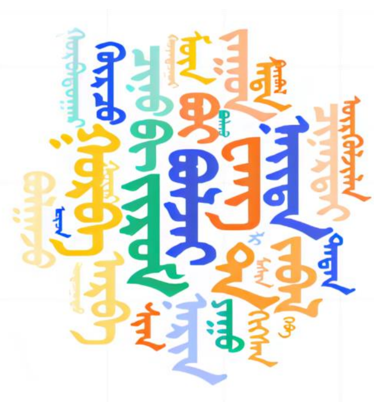

Figure 1. Word Cloud of High-Frequency Terms (>0.75%).

High-frequency Word Exclusion: Words exceeding the 0.75% frequency threshold were excluded to refine topic extraction, reducing the distortion caused by these terms and ensuring a more accurate representation of the text's content.

2.2.3 Low-frequency word list

To mitigate the impact of low-frequency words on Mongolian topic extraction, which often carry limited topic information or include erroneous terms, we implemented a thresholding strategy.

A word frequency threshold of 0.001 was established to exclude infrequent words from the topic model. Only words with a frequency above this threshold were considered for topic modeling, with those below the threshold being discarded. A partial list of such words is provided in Table 1.

Table 1. Partial List of Low-Frequency Words (Word Frequencies<0.001)

Wrong word	Word frequency
[illegible]	0.00011
[illegible]	0.0004
[illegible]	0.0006
[illegible]	0.0009
[illegible]	0.00092

This thresholding approach ensures that the topic extraction process is more accurate and representative, free from the noise introduced by less common words.

2.2.4 Mongolian topic classification

The number of topics, represented as K, needs to be determined prior to training the topic model. Perplexity [15] determines the optimal number of topics by assessing the model's prediction accuracy on text data. A lower perplexity generally indicates improved prediction performance and typically decreases as more latent topics are introduced. For this experiment, we explored a range from 20 to 80 for the topic number K with an increment size of 5. We computed perplexity values for various topic numbers and illustrated their variations in Figure 2.

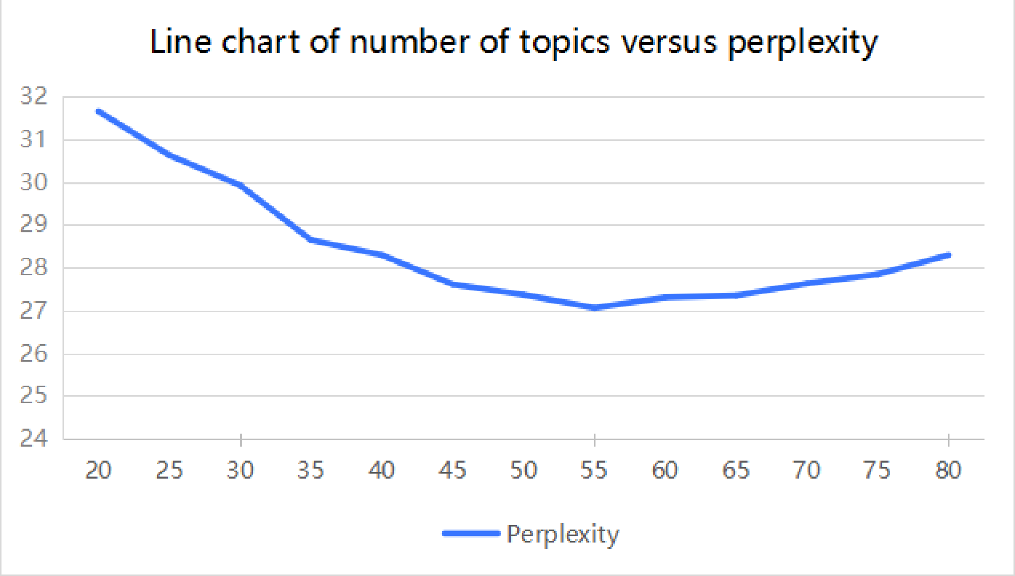

Figure 2. Line chart of topic-perplexity change.

2.3 Pre-trained Mongolian topic model

2.3.1 Introduction of the topic model

The Embedded Topic Model (ETM), introduced by Dieng et al. [6], combines the LDA topic model with Word2vec vectors to capture the interaction between word vectors and topic distributions for document representation. In ETM, Word2vec provides a matrix representation ρ of word vectors, where L is the vector dimension. The document-topic distribution θ yields the kth topic αk as a matrix K×L. ETM generates the word distribution β as a matrix K×V by multiplying the topic vector matrix and word vector matrix. β is used to optimize the alignment between topic and word vectors, enhancing the probabilities of words within each topic k. Figure 3 outlines the architecture of ETM. Fig. 3 outlines the ETM architecture.

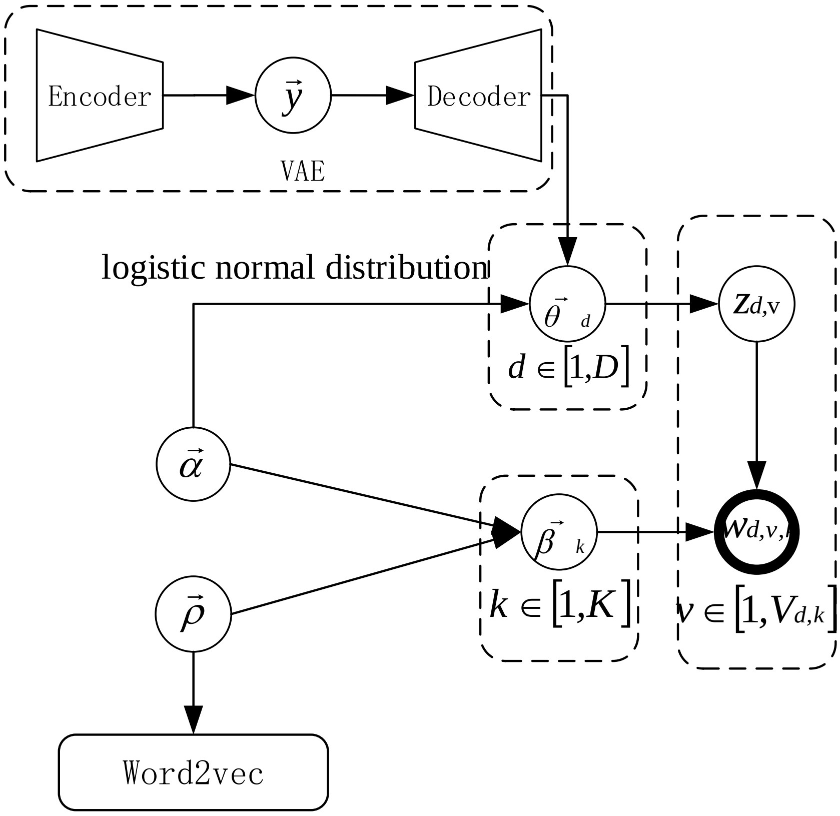

Figure 3. ETM diagram.

2.3.2 Introduction of the pre-trained model

The IMNU-BERT pre-trained model is used in this study to represent Mongolian word vectors. It captures semantic, grammatical, and contextual information about words in Mongolian through pre-training on the Mongolian corpus. This high-quality representation improves the accuracy of Mongolian topic extraction and supports text-processing tasks.

2.3.3 Introduction to Mongolian topic models based on pre-trained models

The challenges in Mongolian topic extraction methods currently stem from limited text data and a lack of model generalizability. Additionally, existing approaches primarily focus on traditional methods for constructing Mongolian topic models. To address these issues and extract Mongolian topics based on semantic understanding, this paper proposes an improved version of the ETM topic model called IMNU-BERT-ETM. The IMNU-BERT-ETM enhances the ETM by dynamically obtaining rich semantic vectors ρ through the use of the IMNU-BERT pre-trained model instead of static word vectors from Word2vec in the ETM. The representation β for topic word distribution is derived by multiplying the semantic vector with the topic vector. Finally, using a constructed Mongolian thesaurus as a reference, the variational autoencoder infers the topic words within texts. Fig. 4 illustrates IMNU-BERT-ETM.

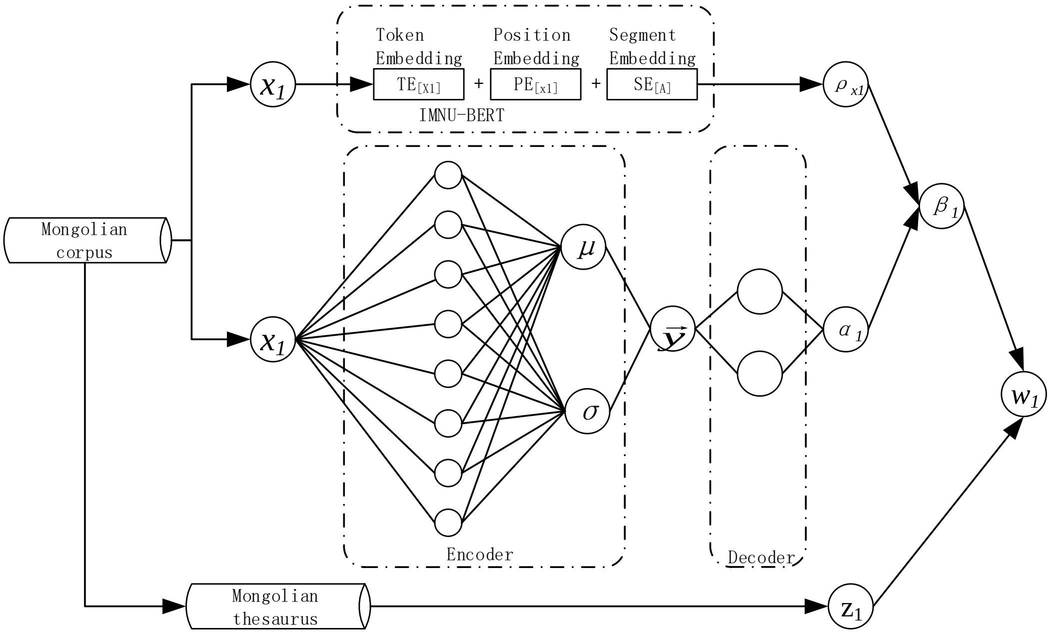

Figure 4. IMNU-BERT-ETM diagram.

The key steps in building an IMNU-BERT-ETM are as follows:

1) Text preprocessing: The Mongolian corpus data undergoes preprocessing, including the removal of stop words and filtering of high-frequency and low-frequency words.
2) Thesaurus construction: The Mongolian thesaurus is constructed by preprocessing Mongolian text.
3) Topic vector representation: The variational autoencoder generates the topic vector representation, while a pre-trained model generates the word vector representation of Mongolian text.
4) Joint training of topic model: Train the topic model jointly using both the topic vector and word vector.
5) Topic word extraction: Obtain the document's topic words based on predictions from the topic model.

The Mongolian topic model based on pre-trained models (IMNU-BERT-ETM) makes full use of the advantages of pre-trained models in semantic understanding and combines them with a topic model to realize unsupervised topic extraction.

3. Experiments and analyses of results

3.1 Experimental design of the IMNU-BERT-ETM

The study validates a Mongolian topic extraction approach using a pre-trained model, comparing it with LDA and ETM. We also evaluate the impact of stop-word removal on the performance of IMNU-BERT-ETM. Our dataset, obtained from the Mongolian Big Data Research Base at Inner Mongolia Normal University, covers various topics such as history, culture, politics, and economics. The experimental method involves several steps: 1) Preprocessing the Mongolian corpus by removing stop words and filtering high-frequency/low-frequency words to enhance data quality for topic extraction; 2) Creating a Mongolian thesaurus by identifying key topic words through text preprocessing to facilitate extraction; 3) Generating topic vector representations using a variational autoencoder for topics and a pre-trained word vector model in Mongolian; 4) Training a topic model with both topic and word vectors to learn latent topic distribution in text; 5) Extracting relevant core content suitable for tasks like analysis and classification based on inferred topics from trained models.

3.2 Experimental parameter settings

The algorithm shows good prediction performance on the text with K=55, which is determined as the optimal value for the topic number. A batch_size of 256 was used for model training, and a variational autoencoder and cross-entropy loss were applied with a learning rate of 1×10^{-3}. The preprocessed dataset of 2.67 million Mongolian text images was divided into training, validation, and test sets at a ratio of 9:0.5:0.5 respectively. Python3.7 and PyTorch deep learning frameworks were utilized to conduct experiments on an NVIDIA GeForce GTX 3090-equipped server.

3.3 Evaluation metrics

This study used the product of topic diversity [6] (TD) and topic coherence [16] (TC) to measure topic quality (Q).

Topic diversity is the percentage of different words appearing in the top N words of all topics. A diversity close to one indicates that the topic is more diverse, and vice versa indicates topic redundancy. The calculation formula is given in Equation (1).

$$TD = \frac{unique(w_{topN}^{1}, w_{topN}^{2}, \ldots, w_{topN}^{k})}{K \times N} \tag{1}$$

Topic coherence is given by Equation (2):

$$TC = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} f\left(w_i^{(k)}, w_j^{(k)}\right) \tag{2}$$

where $\{w_1^{(k)}, \ldots, w_{10}^{(k)}\}$ denotes the top 10 most likely words in topic k. Let $f(w_i,w_j)$ denote the normalized, pointwise, mutual information. Subsequently, topic relevance is measured by the average, pointwise, mutual information between any two words contained in topic k. Its formula is shown in Equation (3). If the co-occurrence probability of a word in topic k is high, the topic is more consistent and interpretable.

$$f(w_i, w_j) = \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-log\, P(w_i,w_j)} \tag{3}$$

where $P(w_i, w_j)$ is the probability of a word w_i and w_j co-occurring with the document and $P(w_i)$ is the marginal probability of w_i.

Q is an indicator that is used to evaluate the overall performance of a topic model. It is a product of topic relevance and diversity. The better the performance of the model is, the closer the result is to one. Equation (4) shows the formula.

$$Q = TC \times TD \tag{4}$$

3.4 Experimental results and analyses

3.4.1 Topic model performance comparison

In the first set of experiments, the LDA, ETM, and IMNU-BERT-ETM models were benchmarked for topic extraction capabilities. Table 2 summarizes the results, highlighting IMNU-BERT-ETM's superiority in topic relevance (TC) and diversity (TD), leading to improved overall topic quality (Q).

Table 2. Topic model performance comparison results

Topic models	TD	TC	Q
LDA	0.9551	0.2825	0.2698
ETM	0.7539	0.7211	0.5429
IMNU-BERT-ETM	0.8012	0.7619	0.6104

The data reveal that IMNU-BERT-ETM achieved a higher TD value than LDA but a lower TD value than ETM, indicating a more focused extraction on core topic words, potentially at the expense of topic generalization. However, the increase in TD compared to ETM suggests that IMNU-BERT-ETM maintains reasonable topic diversity. Moreover, IMNU-BERT-ETM significantly outperformed LDA and ETM in TC, enhancing the capturing of lexical relationships within topics and improving the coherence and interpretability of the extracted topics. These findings demonstrate the superior performance and generalizability of IMNU-BERT-ETM in Mongolian text-processing tasks.

3.4.2 Removing the effect of stop words

The second set of experiments assesses the impact of eliminating stop words on the IMNU-BERT-ETM's performance, as shown in Table 3. The results indicate that removing stop words enhances topic quality, reaffirming the importance of this step in topic modeling to avoid noise and improve accuracy.

Table 3. Comparison of the impact of stop words on topic model performance

Topic Models	Whether to remove stop words	TD	TC	Q
IMNU-BERT-ETM	no	0.8012	0.7619	0.6104
IMNU-BERT-ETM	yes	0.8024	0.7679	0.6162

The experimental data reveal that stop word removal boosts the TD value by a margin of 0.0012, albeit slightly. Nonetheless, this contributes to a richer representation of topics by allowing the model to focus on relevant terms. The TC value increases by 0.0060 when stop words are removed, suggesting a better correlation between words within topics. Additionally, the Q value for the model with removed stop words is higher by 0.0058, signifying an overall improvement in topic quality due to increased relevance and diversity. These findings emphasize the critical role of stop-word removal in enhancing topic modeling outcomes.

3.4.3 Analysis of topic word distribution results

The IMNU-BERT-ETM model efficiently identifies representative words for each topic, as demonstrated by the Mongolian topic word distributions presented in Table 4. This table highlights the five most prevalent topic word distributions across the model's five groups, along with the five words highest in probability for each topic, encompassing domains such as politics, plants, culture, education, and health.

Table 4. IMNU-BERT-ETM topic word distribution

Topic	Word	Field of affiliation
Subject1	[illegible]	political
Subject25	[illegible]	plants
Theme27	[illegible]	culture
Subject37	[illegible]	education
Subject 8	[illegible]	health

The experimental outcomes validate that our pre-trained model-based Mongolian topic extraction approach exhibits superior performance and generalizability on Mongolian text data. It proficiently extracts precise topic information from Mongolian text, offering a robust mechanism for topic discovery in Mongolian text processing tasks.

4. Conclusion

The paper presents a Mongolian topic extraction approach that utilizes a pre-trained model. Initially, Mongolian text undergoes correction, cleaning, and segmentation. Then, a Mongolian thesaurus is created by removing stop words and filtering high-frequency and low-frequency terms. We combine the pre-trained model with a topic model to generate Mongolian word vectors for unsupervised extraction. Experimental results demonstrate that our method outperforms LDA and ETM in terms of accuracy and reasonableness of extracted topics, improving quality by 0.3406 and 0.0675 respectively. This highlights the effectiveness of our proposal.

Future research will explore integrating additional NLP technologies like sentiment analysis and text classification to expand applications for Mongolian topic extraction. Furthermore, we aim to validate the practical usefulness of our method in domains such as intelligent question answering and information retrieval, thereby advancing Mongolian NLP.

Acknowledgments

We thank Professor Wang Siriguleng and Associate Professor Si Qintu for their guidance and help in the research process.

Funding

This work was supported by the Natural Science Foundation of Inner Mongolia (2022MS06002), the Science and Technology Plan Project of Inner Mongolia Autonomous Region (2021GG0139), and the Open Project Foundation of Inner Mongolia Discipline Inspection and Supervision Big Data Laboratory (IMDBD202109).

References

[1] Blei D M, Ng A, Jordan M I. (2003). Latent Dirichlet allocation. J Mach Learn Res., 10:1162. DOI:2003/JMLR 2003.3.4-5.993.

[2] Blei D M. (2011). Probabilistic topic Modelsm Proceedings of the 17th ACM SIGKDD International Conference Tutorials. DOI:10.1145/2107736.2107741.

[3] Das R, Zaheer M, Dyer C. (2015). Gaussian LDA for topic models with word embeddings. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015:795-804.

[4] Yang M, Cui T, Tu W. (2015). Ordering-sensitive and semantic-aware topic modeling. Proceedings of the AAAI Conference on Artificial Intelligence, 29:1.

[5] Miao Y, Grefenstette E, Blunsom P. (2017). Discovering discrete latent topics with neural variational inference. International Conference on Machine Learning. PMLR 2017:2410-2419.

[6] Dieng A B, Ruiz F J R, Blei D M. (2020). Topic modeling in embedding spaces. Trans Assoc Comput Linguist, 8:439-453.

[7] Grootendorst M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. DOI:10.48550/arXiv.2203.05794.

[8] Yuhan W, Min L, Yanling L, et al. (2023). Research on embedded text topic model based on BERT. CEA, 59:01.

[9] Yupeng J. (2016). Research on the topic model of Chinese text based on LDA. Inner Mongolia University.

[10] Riguleng S. (2016). Research and system implementation of Mongolian information retrieval based on LDA. Inner Mongolia Normal University.

[11] Fu Y J. (2018). Mongolian short text semantic similarity calculation based on deep VAE algorithm with topic Information. Inner Mongolia University.

[12] Shuguang B. (2021). Mongolian text keyword extraction and word analysis. Inner Mongolia normal university. DOI:10.27230/dcnki.Gnmsu.2021.000910.

[13] Zheng G, Gaowa G. (2011). Comparative study on Mongolian stop words and English stop words. JCIP 25:35-38.

[14] GB/T 26235-2010. Mongolian word tagging for information technology information processing GB/T 26235-2010.

[15] Zhang Z, Han X, Liu Z, et al. (2019). ERNIE: Enhanced language representation with informative entities. DOI:10.18653/v1/P19-1139.

[16] Mimno D M, Wallach H M, Talley E M, et al. (2011). Optimizing semantic coherence in topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. Association for Computational Linguistics.

[17] Xiaobao W, Thong N, Tuan A L. (2024). A survey on neural topic models: methods, applications, and challenges [J]. Artificial Intelligence Review, 2024, 57(2).