International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 4305-4312

https://doi.org/10.70517/ijhsa463368

Analyzing Nature Imagery in Tang Poetry Using LDA Thematic Modeling

Xiaoyi Dong[1,*]
[1] China University of Petroleum (Beijing) Karamay, Karamay, Xinjiang, 834000, China
Corresponding authors: (e-mail: d2776401085@163.com).

Abstract In this study, a corpus selected from All Tang Poems was collected, and the information in the text about the study of this paper was extracted through data cleaning, de-duplication and other operations. The LDA topic model is used to classify the topics of Tang poems according to the topic words appearing in the poems. Combined with the TF-IDF algorithm, the probability distribution of natural imagery themes in Tang poems is calculated. By analyzing the emotional indexes of different natural imagery themes in Tang poems, the emotional characteristics embodied by the author in the poems are studied. In this paper, the LDA method is used to categorize the themes of Tang poems more accurately, and the theme words contained under each Tang poem theme are closely related. TF-IDF can be used to effectively determine the themes of Tang poems, such as the probability distributions of "natural imagery" and "wandering the world" in "The End of Spring", which are 0.214 and 0.550, respectively. In Tang poetry, the theme of "natural imagery" had the lowest positive affective index of "cold rain" (0.17), and "pine and cypress" had the highest positive affective index of 0.83. In Tang poems, the high-frequency words related to the artistic conception of mountains are "moon", "rain", "smoke", etc., which appear more than 15 times, which verifies the applicability of the LDA model in the text mining of Tang poems.

Index Terms LDA topic model, TF-IDF algorithm, Tang poetry, natural imagery

I. Introduction

Poetry is the conscious landscape, and landscape is the externalized poetry. Poetry is an important carrier of landscape poetry, and its content shows humanistic scenes, landscape environment, historical background and emotional meaning, vividly reflecting the regional natural landscape and poetic atmosphere [1], [2]. The poetic thoughts expressed in the landscape poems of ancient Chinese literati profoundly reflect people's pursuit of poetic dwelling life and thinking about natural intention [3]. As a spiritual carrier, poetry expresses the literati's perception of natural landscape while presenting the landscape in people's spiritual world, thus influencing the camping of the landscape that exists as a material form [4], [5]. Under the influence of poems and songs, the aesthetic experience brought to people by the beauty of nature is further developed and strengthened, which promotes the creation and enhancement of landscape imagery, and gives more poetic meaning to the natural landscape [6]-[8]. Mining the natural intention and poetic spirit depicted in Tang poems and exploring the needs of contemporary ideal natural landscapes is the mission and task given by the idea of ecological civilization [9].

LDA model is a document topic generation model used to mine the hidden topic information in text [10]. Using this text mining model, the relationship between landscape and poetic emotion can be explored, and the poetic landscape characteristics of different landscape nodes can be further studied in depth from a microscopic perspective [11], [12]. The scope of the study can also be expanded to include poems describing nature through the ages, so as to study the development of poetic natural landscape from a more macroscopic point of view, and provide more theoretical references for the creation and development of poetic landscape in natural landscapes [13], [14].

The article takes the LDA topic model as the core and combines it with the TF-IDF algorithm to classify the keywords on the topics of Tang poems and compute the probability distribution of the topics of natural imagery. During the training of the LDA topic model, the number of topics of Tang poems as well as the hyper-parameters are iteratively updated through the E-steps and the M-steps to ensure the quality of the modeling. Using the TF-IDF algorithm, the probability distributions of documents on different topics and each word on each topic are obtained based on the Tang poetry document-topic and topic-word result matrices output from the topic model. We analyze the sentiment index of natural imagery themes in Tang poems, and take the theme of “mountain color” as an example to explore the characteristics of mountain color mood creation in Tang poems.

II. Research on Nature Imagery of Tang Poetry Based on LDA Theme Modeling

The implicit Dirichlet distribution (LDA) topic model [15], [16] is a topic probabilistic generative model that can be mined to extract the implicit semantic topic structure from a large number of textual datasets, which contains the three hierarchical structures of documents, topics, and words, and mainly obtains the distribution weights of the topics in different documents and each word in each topic through the model's inference operations.

II. A.LDA model derivation

The LDA topic model is mainly used to derive the distribution of different topics in a document and the distribution of different words in related topics by iterative updating, so as to reveal and discover the potential topic structure in massive text data. Therefore, this section will give a brief description of the derivation process of LDA topic model.

II. A. 1) Parameter setting

Before LDA carries out model training, the number of topics and hyperparameters need to be set accordingly, and the number of topics K and hyperparameters α and β need to be constantly modified according to the actual situation of modeling during the training process, so as to ensure the effectiveness of modeling. The number of topics K represents the complexity of the model, if the number of topics is too small, the model's ability to describe the data will be limited. But when the number of topics is too high, it will increase the training time of the model and make no sense to the training results of the model. Therefore, in the modeling process, a cross-validation approach is generally adopted, such as selecting different numbers of topics based on confusion, consistency, and manual judgment to test the validity of the modeling, and then determining the choice of the number of topics K through the final accuracy.

II. A. 2) Iterative update process

In the process of iteratively updating the relevant parameters of the LDA topic model, it mainly consists of two steps, the E step (Expectation Step) and the M step (Maximization Step), which are carried out alternately until the model training is completed.

In the E step, the main purpose is to calculate the distributional expectation of document-topic and topic-word, so as to estimate the topic distribution. Specifically, the probability distribution of each word in each topic of each document will be calculated to determine the relationship between each topic and each word, so as to update the document-topic results.

In the M step, the EM algorithm or gradient descent method is mainly used to optimize and update the number likelihood function to maximize it, and then iteratively optimize the distribution parameters of document-topic and topic-word to gradually improve the fit of the model, thus obtaining better results.

In addition, when optimizing and updating the distribution parameters of document-subject and subject-word, the parameters are also combined with the Delicacy prior distribution to control the parameters within a reasonable range, which makes the model more reliable.

II. A. 3) Algorithm flow

LDA topic model is mainly used to infer the implied text topics and reveal the relationship between documents and topics from the given text data by iterative updating. The following is the process of LDA's topic algorithm:

(1) In the modeling initialization phase: set the initial values of the number of topics K and hyperparameters α and β, and randomly initialize the distribution parameters of document-topic and topic-word.

(2) Iterative updating: first calculate the distribution probability of each word in each topic in each document, and update the distribution of document-topic, which can be specified by the formula P(document/word) = P(document/topic) * P(topic/word) to calculate the probability that each word belongs to each topic, and reflects the possibility of each word under different topics, as well as the relevance of topic and word, topic and document. Then based on the results of the calculated probability distribution, the words under each topic are reassigned using sampling method or maximum probability method, etc. Finally, the document-topic and topic-word distribution parameters are updated.

(3) Repeat the steps of iterative updating until the convergence condition is reached.

(4) Finally output the document-topic and topic-word result matrices: i.e., calculate and obtain the probability distribution of each document in different topics and each word in each topic.

Therefore, the process of modeling the LDA topic model of Tang poetry text is referenced as follows:

(1) Sampling from the Dirichlet distribution α to obtain the topic distribution θ of the Tang poetry document i: i.e., to obtain the distribution of each topic in the document i by sampling from the prior distribution α, i.e., to determine a good distributional weighting of each topic in each document.

(2) Sampling from the topic distribution θ_i of the Tang poetry document i to obtain the topic $Z_{i,j}$ of the j th word in the document i : That is, the topic of the j th word of document i is sampled from a polynomial distribution of topics in document i from the distribution of topics in document θ_i .

(3) Sampling from the Dirichlet distribution β to obtain the topic $Z_{i,j}$ corresponding to the word distribution $\eta z_{i,j}$; i.e., sampling a priori distributions β in the topic $Z_{i,j}$ to get the word distribution corresponding to the topic. distribution of words corresponding to the topic, and determine the distribution weight of each word under the topic.

(4) Sampling from the polynomial distribution of words $\eta z_{i,j}$ to obtain the final word $W_{i,j}$: that is, based on the topic $Z_{i,j}$ corresponding to the distribution of words i, j , and in accordance with the polynomial distribution of the related words to the sample, so as to obtain the document i in the first j of the document. The final word distribution at the j position in the document is obtained by sampling the related words according to the polynomial distribution.

II. B.Model evaluation

In this study, the evaluation of LDA topic models will be performed based on perplexity and consistency. Perplexity is a commonly used evaluation metric that reflects the degree of fit of textual data in the model and indicates the uncertainty of the model with respect to the observed textual data. The formula for perplexity is given below:

$$perplexity(D) = \exp(-\frac{\sum \log p(w)}{\sum_{d=1}^{D} Nd}) \tag{1}$$

where (D) is the total number of documents, $p(w)$ refers to the probability of occurrence of each word in the test set, and Nd denotes the total number of words in the d th document.

Consistency is another important metric used to evaluate LDA topic models, which implies word consistency and relatedness across topics generated by the modeling. Specifically, the consistency metric is used to evaluate the similarity between words in the same topic, with higher values indicating better modeling of the topic. By calculating the consistency score of words in a topic, we can evaluate the model's ability to capture semantic consistency. The formula for consistency calculation is as follows:

$$Coherence = \frac{2}{N(N-1)} \sum_{i-2}^{N} \sum_{j=1}^{i-1} Sim(w_i, w_j) \tag{2}$$

where N denotes the number of words in the topic, w_i and w_j denote the i and j th word, and $Sim(w_i, w_j)$ denotes the similarity between the words w_i and w_j , and in general This value ranges between [0, 1].

With the above evaluation indexes and methods, the LDA topic model can be comprehensively evaluated to reveal the model's performance in document topic clustering.

II. C.TF-IDF based subject distribution calculation

Among the traditional topic extraction algorithms, the simplest, classic and still widely used algorithm is the TF-IDF model [17], [18], which can reflect the relative importance of the feature items such as phrases or sentences in one of the documents in the document set. Because of the relative simplicity of the algorithm and its high recall and accuracy, it has been favored by many researchers and has become a commonly used algorithm for calculating weights in topic extraction, information retrieval, text mining, and natural language processing.

In this study, the method is used to screen keywords to remove high-frequency general words and too remote words, and finally extract the keywords with high word frequency as the label of LDA topic model.

Set a certain document set as $D = \{d_1, d_2, ...d_N\}$, N denotes the number of all documents in D. The weight of a given word is computed using the TF-IDF method, and T is the set of feature terms in the document set, $T = \{t_1, t_2, ..., t_k\}$, and K is the number of feature terms. The formula is as follows:

$$\begin{cases} WTF - IDF = TF \times IDF \\ IDF = \log(N / n) \end{cases} \tag{3}$$

TF (Term Frequency) in the above equation represents the frequency of occurrence of the word t_i in the document d_j .TF can be expressed as:

$$TF_{ij} = \frac{m_{ij}}{\sum_{K} m_{ij}} \tag{4}$$

where m_{ij} denotes the number of times the word t_i occurs in the document d_j , and $\sum_{k} m_{ij}$ is the total number of times all the words occur in the document d_j . For TF, if a word occurs very frequently in a certain document, it means that it has a strong ability in recognizing the contextual properties of that document.

IDF (Inverse Document Frequency) represents the inverse document frequency of a document, as shown in Equation (3), n denotes the number of documents in which the word t appears. However, sometimes the word is not in the document set, which leads to zero divisor, so the IDF formula is generally deformed as:

$$IDF = \log \frac{N}{1+n} \tag{5}$$

For IDF, if a word occurs in a very wide range of documents, then the smaller the IDF value, the lower the attribute of the word to recognize the content of the document. Taking the TF-IDF together, if a word appears very frequently in a document, while appearing in only a few documents, then the TF-IDF value is chemicalized to be higher, which means that the word has a high ability to discriminate the document.

II. D.Data set creation

II. D. 1) Sources of corpus

The corpus in this paper comes from All Tang Poetry. The All-Tang Poetry contains the lifetime work of the poets of the Sheng Tang period in China, and it is not difficult to find out that the writing form of the All-Tang Poetry mainly includes the volume number, the poem number, the title, the author, and the content, and the dataset we need is mainly to read the original ancient poems and authors into the txt file.

II. D. 2) Data pre-processing

The process of organizing and cleaning the raw data is called preprocessing of data. Since there is a problem of mismatch between the data type or format of the dataset and the requirements of the algorithm, the data should be preprocessed first before mining the data. As far as data mining is concerned, it is mainly divided into two parts: data preprocessing and data processing. An indispensable step in data mining is to perform data preprocessing. The task of data preprocessing is to provide accurate, effective and targeted data for the data mining algorithm, eliminate those data or attributes that are not relevant to data mining, and correct the format of the data and other operations, so that the format of the data is unified, which can improve the efficiency of data mining, and then improve the accuracy of data mining.

Pre-processing of data, on the one hand, the unified collation of text data is mainly used in the author and content. Among them, the title of the volume (“Volume 56”, “Volume 84”, etc.), copyright information and other content is the invalid content of the poem, mainly with regular expressions and other methods of each line of the initial cleaning, including the removal of spaces, line breaks and annotations of the proofreader.

Removing stop words refers to filtering out the word segmentation results that appear in the stop word list. The text of ancient poems has certain particularities, mainly because the words used in ancient poems are relatively refined, such as "of" and "had" in the particles; Words such as "dang", "due" and pronouns "you", "me", and "he" in prepositions rarely appear in ancient poems. If it does, it is difficult to judge how much information it contains. At the same time, the grammatical structure of ancient poems is also somewhat different from that of modern works, and if some words are blindly deleted, it may have a certain impact on the classification effect. Therefore, in this article, only the punctuation contained in the dataset is removed.

II. D. 3) Form of data sets

In the processed data set, each piece of data contains two fields: “author” and “data”. The “author” field is the author of the record; the “data” field is the text data of the record itself.

III. Thematic Analysis of Nature Imagery in Tang Poetry under LDA Modeling

III. A. Classification of Tang Poetry Themes Based on LDA Models

In order to verify the effectiveness of the LDA method in extracting the theme of Tang Dynasty poems, it was compared with the PLSA method. The dataset is selected from the Tang Dynasty poetry dataset, the experimental environment and configuration settings are Windows10 system, the CPU is i5-8300H, and the graphics card is GTX1050Ti 4GB. Set the number of topics in PLSA and LDA to the optimal number of topics 6, which are "natural imagery", "reading pilgrimage", "nothing to do in the family and country", "wandering the world", "prodigal son's journey", and "wind and dust in the world", and the number of topics in each topic is 50. The classification results of some topics using the PLSA method and the LDA method are shown in Figure 1 and 2, respectively.

Taking the topic category of "natural imagery" as an example, the results of topic extraction using the PLSA method include "sage", "Guanshan", and other subject words that are very related to national themes. The differentiation of subject headings among the subject categories is poor, and the topic classification results are not ideal. Compared with the topic extraction results of LDA, the results of topic extraction using the LDA method are better than those of the PLSA method. From the perspective of time efficiency, due to the large number of poetry documents, the parameters of PLSA method P(z|d) will also increase linearly with the linear increase, and the running time is significantly higher than that of LDA method. In summary, the LDA method is more effective in extracting the theme of Tang Dynasty poems.

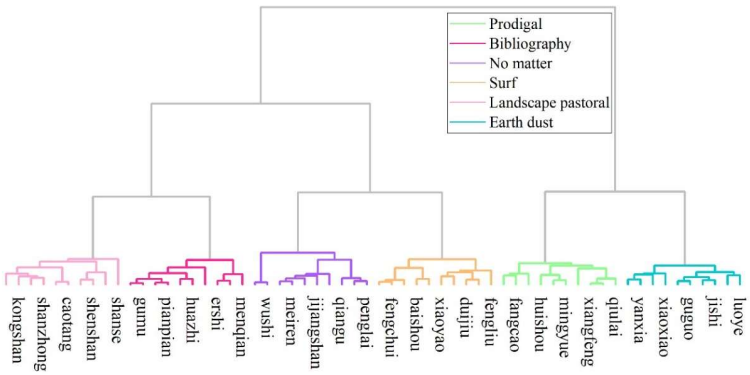

Figure 1: Part of the topic classification results of the LDA method

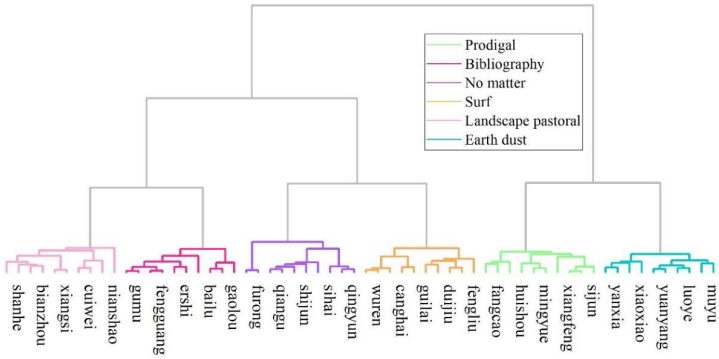

Figure 2: Part of the topic classification results of the PLSA method

III. B. Probability Distribution of Nature Imagery Themes in Tang Poetry

Due to the deep cultural connotation, diversified emotional expression and rich symbolism of ancient poems, they can express different emotions and thoughts through lyricism, argumentation, description and other modes of expression. The imagery and symbols involved often have multiple meanings and derivations, which makes an ancient poem contain multiple themes at the same time. The expression of multiple themes not only enhances the artistry and cultural connotation of ancient poems, but also gives them a wider meaning and value.

Considering the rise and fall of poetic composition and the fact that a short verse may convey multiple thematic tendencies, the probability distribution of themes in Tang poems is calculated with respect to this characteristic, and the method of setting theme thresholds is proposed. When the probability of a theme in a poem is greater than 1/K

(K is the number of themes), the theme is listed as one of the reference themes of the poem. The specific operation process is as follows, take a representative dataset of Tang Dynasty poems as an example, use the TF-IDF method to calculate the topic probability distribution of Tang Dynasty poems and output the results, the results of the topic probability distribution of Tang Dynasty poems are shown in Table 1, and select one of the data for analysis.

Take the first data as an example, according to the set theme threshold 1/K, the calculation can get the themes related to the poem are Topic1 (Natural Imagery) and Topic4 (Wandering to the Ends of the World), and then according to the serial number to look for the poem text dataset, to find its corresponding poem is Li Bai's "Returning to the Pine Niches of Terminal South Mountain in Spring," with the content of the poem "I came to the southern mountain sun, and things are not different from the past. I've come to the south of the mountain, and everything is different from the past. But I seek the water in the stream, and still look at the rocks. The roses edge the east window, and the female roses wrap around the north wall. The grass and trees have grown a few feet. The grass and trees have grown a few feet long. I will order a bottle of wine again, and drink Tao Yongxi alone."

The poem describes in detail the natural imagery of Zhongnan Mountain, which is in line with the theme of landscape and idyll. According to the background of the poem, it can be seen that the poem was written in 731, when Li Bai was 30 years old. He left Anlu and went to Chang'an to seek a political career, but he never got what he wanted, and finally decided to leave and live in Zhennan Mountain. It expresses the poet's proud spirit of defying the power and nobility, and also shows the poet's free character of wandering to the ends of the world. Therefore, according to the TF-IDF method to calculate the theme probability distribution of the content of the poem and set the theme threshold 1/K, it can effectively determine multiple themes of the poem.

Table 1: Theme model and poetic theme modeling

Number	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	0.214	0.054	0.045	0.550	0.057	0.043
2	0.045	0.061	0.056	0.046	0.043	0.710
3	0.072	0.055	0.057	0.575	0.057	0.067
4	0.046	0.035	0.550	0.045	0.050	0.181
…	…	…	…	…	…	…
33841	0.063	0.061	0.614	0.057	0.239	0.062

III. C. Emotional Index Analysis of Nature Imagery in Tang Poetry

In order to more intuitively feel the different emotional impacts of natural imagery on poets in the Tang Dynasty, the verses describing natural imagery with emotional colors in the Tang poetry documents collected in this paper were summarized and analyzed with big data, and the emotional indexes of some natural imagery in Tang poetry were obtained as shown in Figure 3.

The natural imagery in Tang poems is rich and diverse, and this study investigates ten of the most typical images, which are the bright moon, mountain color, flowing water, autumn wind, lonely clouds, falling flowers, cold rain, geese, willow, and pines and cypresses.

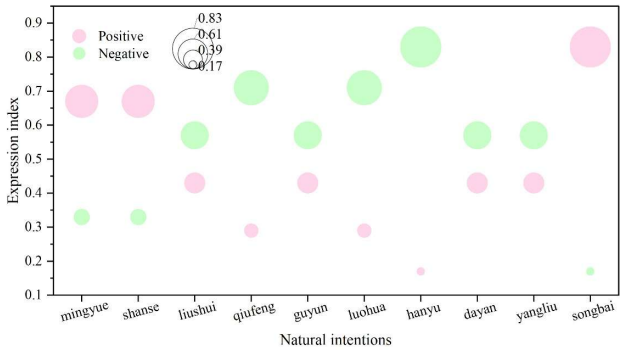

Figure 3: Analysis of the emotional index of tang poetry's natural image

As can be seen from the figure, cold rain has the lowest positive emotional index, only 0.17, followed by falling flowers, only 0.29, indicating that the natural imagery of cold rain and falling flowers is relatively more negatively characterized. Poets in the Tang Dynasty depicted the cold rain as more of a lament for loneliness and nostalgia, reinforcing the poet's inner desolation and helplessness through the coldness and length of the rain. For example,

"You ask for a return date without a date, Bashan night rain rises in the autumn pool", depicting the author's loneliness and the mood of missing his family. Falling flowers embody the images of things being different and life being short, so poets' depictions of falling flowers are mostly characterized by the negative emotions of sadness and loneliness. Pines and cypresses have the highest positive emotion index of 0.83, followed by the bright moon and mountain colors, both of which are 0.67, indicating that the poet shows more positive emotions in the face of these landscapes and the pursuit of natural beauty. For example, "the moon shines between the pines and the clear spring flows over the stones", the pines and the moon constitute a noble mood, reflecting the poet's transcendence.

III. D. Exploring the Characteristics of Mountain Color Context Creation in Tang Poetry

Based on the above macroscopic knowledge of the tendency of emotional index in the natural imagery of Tang poems, the next step is to take the color of mountains as an example of the characteristics of mood creation to be explored. By means of big data analysis, the screened relevant Tang poems are analyzed line by line in terms of word division and semantic data, and the landscape mood characteristics that the mountain color verses have are sorted out through these imagery semantic visualization network analysis diagrams and the statistics of high-frequency words and high-frequency characters.

Fig. 4 shows the analysis of the semantic network (partial) of the artistic conception of Zhongshan in Tang poetry. After data analysis, the landscape images that are more related to "mountain color" include "Yishui", "Nanshan", "Peaks", "Cloud Mountain", "Flying Birds", etc., which reflect the magnificent characteristics of mountain landscape. As a result, there are more related feelings such as "leisurely", "curling", "imagination", "returning to the heart", and "deep valley", which reflect the poet's intoxicated beauty of mountains and rivers, away from the hustle and bustle of life, and the artistic conception of pursuing seclusion.

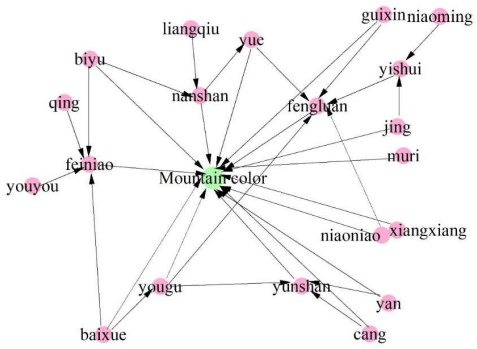

Figure 4: The semantic network analysis of the artistic conception of tang poetry

Figure 5 shows the statistical results of high-frequency words in the artistic conception of mountains in Tang poems, and it can be seen from the analysis data of Tang poems that the mountains in Tang poems mostly present the artistic conception characteristics of "silence", "ethereal" and "clear". The words "cloud", "clear", "moon", "rain" and "smoke" are used more than 15 times. And the relationship with the natural scenery is very close, "cool autumn", "twilight", "wind" and other natural meteorological climate for the Tang Dynasty poems to add more meaning. "Qingshan is far from the ground, and the torrent is high from the sky." The poet is intoxicated by the confrontation of mountains and the magnificent landscape of water flowing down from the heights, expressing the transcendent leisure and the pursuit of the ideal state in his heart.

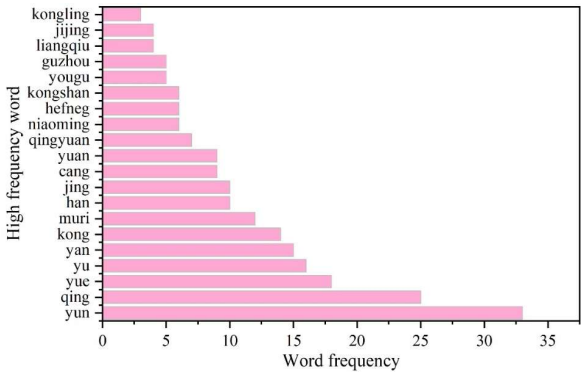

Figure 5: The high frequency word statistics of mountain color imagery

IV. Conclusion

In this paper, we adopt the implicit Dirichlet distribution (LDA) theme model to quantitatively analyze and theme mine the natural imagery in Tang poems.The LDA model improves the effectiveness of theme extraction in Tang poems by iteratively updating the parameter settings. The document-topic and topic-word result matrices of Tang poems are outputted, and the TF-IDF algorithm is utilized to calculate the distribution probability of natural imagery topics in Tang poems. By statistically analyzing the emotion index of natural imagery in Tang poems, we analyze the different feelings of natural imagery to the poets, and analyze the theme of “mountain color” in detail. The following results are obtained:

(1) The accuracy of the LDA method for the topic extraction of Tang poems is better than that of the PLSA method, and the classification speed is faster.

(2) The probability distributions of poem 1 on the themes of "natural imagery" and "wandering the world" are 0.214 and 0.550, respectively, and the corresponding poem is "The Old Hidden Pine Shrine in Nanshan Mountain" by Li Bai.

(3) In the theme of "natural imagery" in Tang poetry, poets often use "cold rain" and "pine and cypress" to convey negative and positive emotions respectively, and the positive emotional indices of the two are 0.17 and 0.83, respectively.

(4) "Cloud", "clear", "moon", "rain" and "smoke" appear more often in the artistic conception of the mountain, and the frequency of words exceeds 15 times.

References

[1] Mehl, K., Gugliano, M., & Belfi, A. M. (2023). The role of imagery and emotion in the aesthetic appeal of music, poetry, and paintings. Psychology of Aesthetics, Creativity, and the Arts.

[2] Maharana, S., & Kaur, Z. (2018). THE POETIC VISION IN THE LANDSCAPE IMAGERIES OF DARUWALLA. Literary Endeavour, 9(4).

[3] Wang, Y. (2024). Research on Natural Imagery and Emotional Expression in Classical Chinese Poetry. Journal of Art, Culture and Philosophical Studies, 1(2).

[4] Wang, Q. (2017). The expressive forms of natural imagery in Chinese poetry. Advances in Literary Study, 5(01), 17.

[5] Feng, Z., Wang, W., & Cupchik, G. (2024). Chinese esthetics through language in poetry: A comparative study of the Chinese wényán and modern báihuà. Arts & Communication, 2(1), 1825.

[6] Nabiloo, A., Baghi, M., & Niazi, F. (2023). The Study of Images of Nature Elements in the Poetry of Children and Adolescent Poets. Journal Of Linguistic and Rhetorical Studies, 14(31), 349-378.

[7] QIN, N., & FAN, X. T. (2022). “Admiration of Flowers in Poems”—On Image Rendering of Flowers in Tang Poems Under the Guidance of Green Translation. Journal of Literature and Art Studies, 12(3), 233-238.

[8] Zhong, S., Peng, H., Li, P., & Xiao, X. (2024). Poetry and the tourist being-in-the-world: connotations behind the Tang Poetry. Current Issues in Tourism, 27(9), 1421-1440.

[9] Zorkina, M. (2018). Describing Objects in Tang Dynasty Poetic Language: A Study Based on Word Embeddings. Journal of Chinese Literature and Culture, 5(2), 250-275.

[10] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia tools and applications, 78, 15169-15211.

[11] Hao, X., Ge, S., Zhang, Y., Dai, Y., Yan, P., & Li, B. (2020). The construction and analysis of annotated imagery corpus of three hundred tang poems. In Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20 (pp. 517-524). Springer International Publishing.

[12] Qiao, J., Xi, X., Zhang, G., & Liang, S. (2024). Interpretation of associative cultural landscape based on text mining of poetry: taking Tianmu Mountain on the Road of Tang Poetry in Eastern Zhejiang as an example. Heritage Science, 12(1), 21.

[13] Xi, X., An, X., Zhang, G., & Liang, S. (2022). Spatial patterns, causes and characteristics of the cultural landscape of the Road of Tang Poetry based on text mining: take the Road of Tang Poetry in Eastern Zhejiang as an example. Heritage Science, 10(1), 129.

[14] Xiao, W. U., Xin, L. I., & Wei, Z. H. A. O. (2019). Research on cultural landscape patterns of Guanzhong area based on text mining of Tang poetry. Landscape Architecture, 26(12), 52-57.

[15] Walaa AlKhader,Khaled Salah,Ahmad Mayyas & Mohammed Omar. (2025). Hydrogen economy research using Latent Dirichlet Allocation topic modeling: Review, trends and future directions. Cleaner Engineering and Technology,26,100953-100953.

[16] Qamar Muneer & Muhammad Asif Khan. (2025). Role of YouTube in creating awareness of sustainable transportation: A Latent Dirichlet Allocation approach. Sustainable Futures,9,100607-100607.

[17] Emre Delibaş. (2025). Efficient TF-IDF method for alignment-free DNA sequence similarity analysis.. Journal of molecular graphics & modelling,137,109011.

[18] Seonghyun Park,Seungmin Oh & Woncheol Park. (2025). Automated Classification Model for Elementary Mathematics Diagnostic Assessment Data Based on TF-IDF and XGBoost. Applied Sciences,15(7),3764-3764.