Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Analysis of Natural Imagery in Tang Poetry Texts Based on Principal Component Modeling

Xiaoyi Dong[1,✉]

[1] *China University of Petroleum (Beijing) Karamay, Karamay, Xinjiang, 834000, China*

ABSTRACT

Tang poetry, as a treasure of ancient Chinese literature, contains a wealth of natural imagery, which not only add to the picture sense of Tang poetry, but are also important carriers of the poet's emotions and thoughts. The study outlines the nature imagery from the perspective of Tang poetry, as well as the key elements and intrinsic connections among them, and borrows k-means clustering to categorize the nature imagery groups. In addition, the study improves the principal component model by using index homogenization, homogenization, and entropy weighting, so that it achieves the best dimensionality reduction effect while guaranteeing the integrity of the data of Tang poetry text.The F1 value of SVM and KNN classifiers for classifying the natural imagery and emotional expression of Tang poetry text is more than 0.9 after dimensionality reduction of the method in this paper, which is a good classification performance. Cluster analysis divides the natural imagery of Tang poetry into astronomical imagery, landscape imagery, and animal imagery, which account for 38%, 53%, and 9%, respectively. “Old times - bright moon”, “Thinking - slanting sun”, “Looking back - west wind”, “the end of the world - west wind” natural discourse is more likely to form word clusters in the natural imagery of Tang poetry. The analysis of principal component model shows that poets are more willing to express their emotions through natural imagery, and the proportion of neutral emotional expression is 5.17% to 7.43%.

Keywords: k-means clustering; principal component model; homogenization; entropy weight method; nature imagery analysis

1. Introduction

Tang poetry is a treasure of ancient Chinese literature, which incorporates rich natural imagery and deep cultural connotations. Natural imagery is a common form of expression in Tang poetry, which expresses the poet's praise for the beauty of nature and his thoughts on life through the depiction of natural scenery such as mountains, water, flowers and grass [1-3]. At the same time, Tang poems also

✉ Corresponding author.
E-mail address: d2776401085@163.com (X. Dong).
Received 25 July 2024; Revised 10 October 2024; Accepted 05 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-068

incorporate cultural connotations such as family, friendship, love and hometown, showing people's deep feelings for family, friendship, love and hometown. This combination and sublimation of natural imagery and cultural connotations not only inherits the excellent traditional Chinese culture, but also influences the aesthetic concepts and literary creation of modern people [4-6].

Imagery is the concentration of poetic qualities, it can be said that poetry is composed of imagery, it is not appropriate to say that poetry is composed of objects, or feelings, or words. Objects point to the objective world, feelings point to the subjective world, words point to the symbolic world, and the three are fused together to become poetic imagery [7-9]. Ancient poems contain many images, all of which are based on natural and social objects. According to the source of imagery, the source of natural imagery can be various things in nature, including but not limited to astronomical phenomena such as the sky, sun, moon, stars, clouds, and rainbows, geographical landscapes such as mountains, rivers, lakes, oceans, and islands, vegetation landscapes such as flowers, trees, grasslands, and forests, biological landscapes such as animals, insects, and fish, and natural phenomena such as natural disasters and climate change [10-13]. All of these natural scenes and phenomena can be a source of inspiration for poets, painters, writers and other artists, as well as a source of beautiful memories and emotional support in people's daily lives [14-15].

In classical poetry, natural imagery occupies a prominent position. As one of the most life-affirming imagery, natural imagery is the source of material for ancient poems, and an important carrier for ancient literati to express their feelings and aspirations [16-17]. Unable to find spiritual relief from the backward productivity, ancient literati could only feel the same rhythm of life as themselves in natural labor, reflect on their integration with life, and realize their interest in natural objects [18-19]. Their descriptions of natural objects such as flowers, plants, trees, mountains, rivers, sun and moon made an objectively existing object become the natural imagery emotionalized and aestheticized by ancient literati, and these natural images have become the symbols of their life experience [20-21]. As an excellent traditional Chinese culture, learning classical poetry not only helps to improve students' language expression ability, aesthetic creativity and thinking development, but also helps to promote students' cultural self-confidence, identification with Chinese culture, love of Chinese culture, and spiritual nourishment [22]. The natural scenery such as mountains and rivers, flowers and trees in nature can make ancient writers feel pleasure, in addition, the change of seasons in nature can also make ancient writers realize the flourishing and declining of life. Ancient writers' strong sense of life in the spiritual world was also an important reason for the formation of natural imagery in ancient poems. The formation of natural imagery is that ancient literati experienced the interest and attitude of life from external objects, and through empathy, it became the object by which ancient literati expressed their emotions [23-24].

The study collects the original ancient poems and authors from All Tang Poetry, and finally builds the Tang poetry dataset for the analysis of natural imagery in this article through data preprocessing (cleaning, word splitting, etc.). For the analysis of natural imagery in Tang poems, the article adopts the improved principal component model to downscale the textual data of Tang poems to achieve more accurate natural imagery with high-frequency noun word frequency statistics and sentiment analysis. By analyzing the key elements in the natural imagery of Tang poems and the intrinsic connections among them, the article provides a classification basis for the utilization of the k-means algorithm to realize the division of natural imagery groups.

2. Basic theory and key methodology

2.1. Nature Imagery and Classification from the Perspective of Tang Poetry

Classical Tang poetry is a bright star in the treasure house of Chinese culture, which has always been valued for its unique charm and profound cultural connotation since its birth. In order to appreciate poetry, we must first have a basic understanding of “imagery”, which is the basic unit of poetry. In this

paper, we will sort out and define “natural imagery” to pave the way for the subsequent discussion.

2.1.1. Nature Imagery. The natural imagery mentioned in this article refers specifically to all the things that poets take from nature and are not artificially formed, such as wind, rain, clouds, flowers, grass, trees, birds and beasts, etc. In classical Chinese Tang poetry, natural imagery has always been one of the most frequently used groups. In Chinese classical Tang poetry, natural imagery has always been one of the most frequently used groups of imagery, and poets often like to use this imagery, which is different from the secular world, to express their feelings, to express their aspirations, or to praise other people and other things by using a certain imagery.

The causes of nature imagery also have an important influence on its concept. From a historical point of view, ancient China has always been an agrarian society, and the people have been living a self-sufficient life as small farmers. Under such historical conditions, what people had the most contact with was the land and the plants and animals growing on the land. When composing Tang poems, poets would start with the things they were most familiar with when choosing imagery, so as to send out their own emotions.

From an ideological point of view, the people in ancient times had a limited level of understanding, and the powerful forces of nature made people feel fearful, so they would “mystify” nature. The image thinking of the Chinese people also makes people more inclined to recognize those marvelous phenomena and things from the image point of view first. Since the beginning of hieroglyphs, people have been keen to record natural things, such as "sun", "moon", "mountain" and other characters that refer to natural scenery occupy a large part of the hieroglyphs. When language and writing continued to develop in the late Tang Dynasty, and literary forms such as poetry and fugue and novels and biographies appeared, people still preferred to create works around natural things, and “natural imagery” was accompanied by the gradual development and maturity of the Chinese mind.

2.1.2. Imagery elements and their intrinsic connection:

1) Elements of Natural Imagery in Tang Poetry. Objective objects are important elements in the imagery of Tang poems, and what kind of objects the poet chooses to include in the poem and what kind of emotions the objects in the poem represent are closely related to the poet's personal aesthetic and emotional tendencies.

2) Intrinsic connection between elements. The combination of different natural imagery can deepen the depth of the mood created by Tang poetry, making it more infectious and representative. For example, in the two sentences of "seven bridges lead to the blue marsh, and two trees connect the flower pond", the combination of the four scenes of "bridge", "tree", "flower" and "pond" vividly depicts the flourishing flowers and trees by the lake, and the beautiful garden scenery connected by the trestle bridge along the coast.

2.1.3. Nature imagery classification based on k-means algorithm. The basic steps of k-means clustering algorithm [25] to categorize the natural imagery of Tang poetry texts are as follows:

1) In the Tang poetry text dataset X, randomly select k data objects as initial clustering centers.

2) Categorize the data objects, calculate the distance between each data object and the k clustering centers, select the clustering center with the closest distance, and classify the data objects into that class, and so on, and classify each data object into the class corresponding to the clustering center with the closest distance.

3) Based on the results, calculate the arithmetic mean of all data objects in each cluster and reselect the clustering center.

4) Re-divide the data objects into classes according to the new clustering centers.

5) Re-step (4) and terminate the clustering if the clustering center no longer changes significantly.

6) Output the results.

2.2. Improved Principal Component Modeling for Nature Imagery Analysis

2.2.1. Basic Principal Component Modeling. The central idea of Principal Component Analysis [26] is to reduce the dimensionality of a data set consisting of a large number of correlated variables while retaining as much information as possible about the variables present in the data set. By transforming into a new set of variables that are uncorrelated and ordered, the first few variables retain most of the variation present in all the original data.

Principal component analysis is an effective method of data dimensionality reduction, which can build more accurate predictive models, thus achieving better dimensionality reduction by optimizing the principal component analysis.

The basic idea of principal component analysis method is to transform the dimensionality, through orthogonal changes, convert the correlated variables into new uncorrelated variables, convert the covariance matrix of the original data variables into diagonal matrix, convert the original data variables into a new orthogonal system, and reduce the dimensionality through multidimensional variables.

Assuming, the distribution direction of the sample data is not axis x_1 and x_2, its distribution trend has a certain regularity, so considering that there may be some kind of connection between the two, it is possible to replace x_1 and x_2 with other variables. However, if any one of these dimensions is used as a substitute, it will definitely cause the loss of the original data information. Therefore, the coordinate system is rotated counterclockwise by a certain angle Θ, so that the direction of the sample points with the greatest degree of dispersion is the z_1-axis, and its orthogonal direction is the z_2-axis, as in the z_1 and z_2 rotation formula is:

$$\begin{pmatrix} Z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\Theta \\ -\sin\Theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = P^T x \tag{1}$$

where P^T is the orthogonal matrix.

The calculation steps of the principal component analysis method are as follows:

1) Standardize the raw data and unify the data scale and order of magnitude.
2) Find the covariance matrix.
3) Find the eigenvalues of the covariance matrix and calculate the variance contribution rate as well as the cumulative variance contribution rate.
4) Determine the number of principal components and derive the results.

2.2.2. Shortcomings in the standardization of principal component models. If the raw data is normalized:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x_j}}{s_j}, i = 1, 2, \ldots, m; j = 1, 2, \ldots, n \tag{2}$$

where $\overline{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$, $s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \overline{x_j}\right)^2$, and $j = 1, 2, \ldots, n$.

While eliminating the effect of the raw data's magnitude and order of magnitude, it also eliminates the information about the differences in the degree of variability of each indicator. The raw data includes two parts of information: one part is the information on the variation of each indicator. The other part is the correlation information between the indicators. As a result of standardization, the variance of each indicator is 1, thus eliminating the differences in the degree of variability of each indicator, but the principal components extracted from the standardized data can only derive one piece of information about the interaction between the indicators, which cannot accurately reflect the information contained in the original data.

2.2.3. Indicator homogenization and indicator convergence. The traditional principal component analysis eliminates not only the effect of the original data magnitude and order of magnitude, but also the variability among the information when standardized. Therefore, the traditional principal component analysis is improved.

Assuming that there is m object and n indicators, the original indicators can form matrix $X_{ij}=\left(x_j\right)_{m\times n}$, and the original indicators divided by the average of all indicators form Z_{ij}:

$$Z_{\bar{j}}=x_{\bar{j}}\sqrt{x_j},i=1,2,\ldots,m;j=1,2,\ldots,n \tag{3}$$

where $\overline{x_j}=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$ and $j=1,2,\ldots,n$.

To make all the indicators act in the same direction on the whole, it is necessary to carry out the process of convergence of indicators, assuming that z_{ij} in Z_{ij} is an inverse indicator, $\min_{1\le i\le m}\left\{z_{ij}\right\}$ is the minimum value, and the index convergence is:

$$z'_{ij}=z_{ij}-\min_{1\le i\le m}\left\{z_{ij}\right\},i=1,2,\ldots,m;j=1,2,\ldots,n \tag{4}$$

z_{ij} convergence processed to z'_{ij}, which does not change the distribution of the original indicator, but also can achieve the dimension reduction.

2.2.4. Entropy weight method. The entropy weight method [27] determines the weights according to the size of the differences of the indicators, which is suitable for the use of more indicators and based on objective data information. In this paper, the entropy weight method is used to improve the traditional principal component analysis method.

The principle of entropy weight method is as follows:

1) Raw data normalization. Perform a normalization transformation on $X_{ij}=\begin{bmatrix}x_{11}&x_{12}&\ldots&x_{1m}\\x_{21}&x_{22}&\ldots&x_{2m}\\\ldots&\ldots&\ldots&\ldots\\x_{n1}&x_{n2}&\ldots&x_{nm}\end{bmatrix}$:

$$y_{ij}=\frac{x_{ij}-\overline{x_j}}{s_j} \tag{5}$$

Where $\overline{x_j}$ is the sample mean of the jnd $n$indicator, $\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$. s_j is the sample standard deviation of the j th indicator $s_j^2=\frac{1}{n-1}\sum_{i=1}\left(x_{ij}-\overline{x_j}\right)^2$, let y_{ij} be the sample data after standardization, the original data matrix X_{ij} becomes $Y_{ij}=\begin{bmatrix}y_{11}&y_{12}&\ldots&y_{1m}\\y_{21}&y_{22}&\ldots&y_{2m}\\\ldots&\ldots&\ldots&\ldots\\y_{n1}&y_{n2}&\ldots&y_{nm}\end{bmatrix}$ after standardization.

2) Standardize the standardized data. In the utilization process, if the indicators contain negative values, the entropy weighting method for weights is invalid, so the use of standardized data to eliminate negative values, without deleting the indicator data to ensure the integrity of the data:

$$y'_{ij}=y_{ij}+\max\left|\left[y_{ij}\right]\right| \tag{6}$$

where $\left|\left[y_{ij}\right]\right|$ denotes the absolute value taken for the integer part of the data.

3) i rd sample y for the j nd indicator, share:

$$p_{ij}=\frac{y'_{ij}}{\sum_{i=1}^{n}y'_{ij}} \tag{7}$$

4) Information entropy of indicator x_j of item j:

$$H_j=-\frac{1}{\ln(n)}\sum_{i=1}^{n}p_{ij}\ln\left(p_{ij}\right) \tag{8}$$

5) Entropy weights for indicator x_j of item j:

$$v_j = \frac{1 - H_j}{\sum_{j=1}^{m}\left(1 - H_j\right)} \tag{9}$$

2.2.5. Improvement of the principal component model analysis algorithm. The improved PCA model divides the Tang poetry text into several text sub-blocks of the same size, which can better display the local information of the whole text.

The training sample consists of L Tang poem texts of size $m \times n$. Each text is divided into $p \times q$ sub-blocks. Matrix A_i is shown in the following equation:

$$A_i = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ A_{p1} & A_{p2} & \cdots & A_{pq} \end{bmatrix} \tag{10}$$

where training sample $(A_i)_{kl}(i = 1, 2, ..., L)$ is partitioned into $p \times q$ new datasets. $\overline{A}_{kl}$ represents the average of all sub-blocks at position (kl) as follows:

$$\overline{A}_{kl} = \frac{1}{L}\sum_{i=1}^{L}(A_i)_{kl} \tag{11}$$

By improving the principal component analysis algorithm the overall distribution matrix of the subtext can be obtained as follows:

$$\begin{aligned} G_{kl} &= \frac{1}{L}\sum_{i=1}^{L}((A_i)_{kl} - \overline{A}_{kl})^T((A_i)_{kl} - \overline{A}_{kl}) \\ \tilde{G}_{kl} &= \frac{1}{L}\sum_{i=1}^{L}((A_i)_{kl} - \overline{A}_{kl})((A_i)_{kl} - \overline{A}_{kl})^T \end{aligned} \tag{12}$$

After calculating the eigenvector $X_{kl}^{(i)}(i = 1, 2, 3, ..., r)$ corresponding to the r largest eigenvalues of G_{kl}, the projection space is obtained $X_{kl} = [X_{kl}^{(1)}, X_{kl}^{(2)}, ..., X_{kl}^{(r)}]$. After calculating the eigenvector $Z_{kl}^{(i)}(i = 1, 2, 3, ..., r)$ corresponding to the d largest eigenvalues of $\tilde{G}_{kl}$, the projection matrix $Z_{kl} = [Z_{kl}^{(1)}, Z_{kl}^{(2)}, ..., Z_{kl}^{(r)}]$ is obtained, and then the eigenmatrix of the Tang poem text.

2.3. Tang Poetry Text Dataset Creation

A corpus is a general term for a storehouse of linguistic materials. With the continuous development of computers and the decreasing cost of computerized storage devices, many researchers have created small corpora that are suitable for their own research directions. And what is used in this paper belongs to the small corpus of personal research content.

This section mainly takes the data collection and data processing part as the starting point, and gives a detailed introduction to the construction process of the Tang poetry dataset, as well as the construction process of the Tang poetry dataset, the source of the corpus, the data preprocessing, and the analysis of the text. The construction process of the specific dataset is shown in Figure 1.

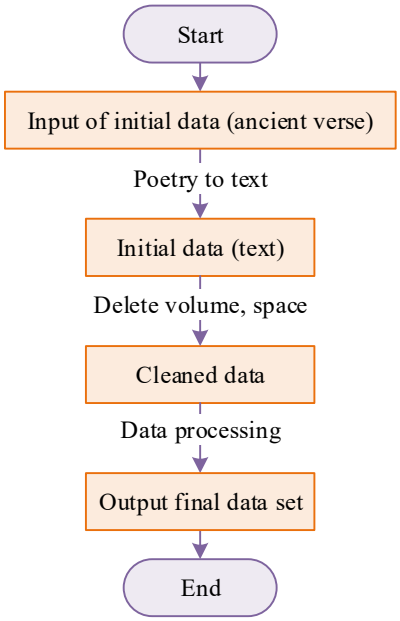

Fig. 1. Flow chart of the dataset construction

2.3.1. Text information extraction:

1) Natural Imagery Word Frequency Analysis. Because the text of Tang poetry belongs to ancient Chinese literature, it cannot be analyzed by simple word division according to ordinary modern Chinese, and it is necessary to carry out certain pre-processing of the Tang poetry material library. Some words with metaphorical meanings are replaced semantically, and some landscape names appearing in Tang poems are replaced by modern recognized name standards, and a vocabulary list of keywords is established on the rostcm6 Tang poetry analysis software for analyzing the word frequency and lexical frequency of Tang poems' landscapes. The whole library of Tang poems describing natural imagery was imported into Rostcm6 for word frequency analysis, and then imported into Excel to create a word frequency table for natural imagery, so that the number of times specific natural imagery was chanted could be obtained.

2) Combination of natural imagery. Imagery itself is ambiguous, and it is difficult for a single image to express the poet's specific inner emotions, so the imagery of each Tang poem has experienced a systematic and orderly combination. Different combinations of imagery produce different spatial moods and convey different emotional effects. The connotation of each imagery symbol is multi-level, such as the landscape imagery of “willow”, when it appears alone, it can be said that it is full of vigor and bright, or it can be said that it is full of longing and sadness, only by placing it in the context of the combination of multiple imagery, can we dig out the real emotion expressed by it.

Therefore, when analyzing the imagery of Tang poems, in addition to focusing on the frequency of the imagery, it is more important to explore the combination relationship between the imagery.

2.3.2. Sources of corpus. The corpus in this paper comes from All Tang Poetry. The All-Tang Poetry contains the lifetime work of Chinese poets in the Sheng Tang period. Looking through the All-Tang Poetry, it is not difficult to find out that its writing form mainly includes volume number, poem number, title, author, and content, and the dataset we need is mainly the original ancient poems and authors, which is read into the txt file. The title of the volume, copyright information and other contents in All Tang Poems are invalid contents of the dataset, which needs to clean the data.

2.3.3. Data pre-processing. The process of organizing and cleaning the raw data is called preprocessing of data [28]. Preprocessing of data, on the one hand, the unified organization of the text

data we mainly use the author and content. Among them, the title of the volume, copyright information and other content is the invalid content of the poem, mainly using regular expressions and other methods to carry out preliminary cleaning of each line, including the removal of spaces, line breaks and the annotation of the proofreader. The grammatical structure of ancient Tang poems is also somewhat different from modern works, so if some words are blindly removed, it may have some effect on the classification effect. Therefore, in this paper, only the punctuation marks contained in the dataset are removed.

2.3.4. Text analysis. In this paper, the types of poems in the dataset and the average number of words were counted. The average length of the poems in this dataset is 45 words, and the distribution of the poems is relatively uniform.

In this paper, we mainly use Python Chinese word processing package "jieba" for word processing. jieba supports three modes: full mode, precise mode and search engine mode. Among them, the full mode is mainly to cut out all the possible words in the words, and this mode is faster, but the shortcoming of this mode is the existence of redundant data. Precise mode tries to slice the complete utterance precisely, the advantage over full mode is that there is no redundant data. The search engine mode is based on the precise mode, and the long words of the precise mode are again sliced, and this participle mode can improve the recall rate. Nowadays, the expression of literature is mainly based on words, but in ancient poems, words are used less, usually words are words. So in this paper, the search engine mode is first selected to split the content of ancient poems, and then the whole poem is split word by word.

3. Results and analysis

3.1. Classification performance of natural imagery feature words

In order to compare with the optimized Principal Component Model (PCA) method proposed in this paper, the traditional feature dimensionality reduction methods DF, IG, and CHI methods are selected for experiments in this section. In order to prove the reliability of the results, 2 classifiers are successively chosen for the experiments, namely SVM classification and K Nearest Neighbor (KNN) classification.

Data cleaning and segmentation were performed for 5000 training samples, and a total of 98,563 feature words related to natural imagery were extracted. In this section, a total of experiments with different feature dimensions are designed, and the feature dimensions are taken as 2000, 2400, 3000, 3600, 4200, and 5000, respectively. In the comparison experiments, the selection of parameters of the classifier is always the same, and in order to globally evaluate the classification effect of the category of Tang Poetry Natural Imagery, the evaluation metrics are used to be the F1 value.

Figure 2 demonstrates the experimental results of DF, IG, CHI, and PCA methods combining the two classifiers under different feature dimensions. As can be seen from the figure, it can be seen that the experimental results of the PCA method are overall better than the other three methods under different feature dimensions. In the figure, when SVM classifier is used, the F1 value of PCA method is on average 7.03%, 3.93% and 1.61% higher compared to DF, IG and CHI methods respectively. In addition, when using the KNN classifier, the F1 values of the PCA method are on average 5.52%, 2.23% and 1.01% higher compared to the DF, IG and CHI methods, respectively. Therefore, the features extracted by the PCA method in this paper have improved classification performance for both of the 2 classifiers used.

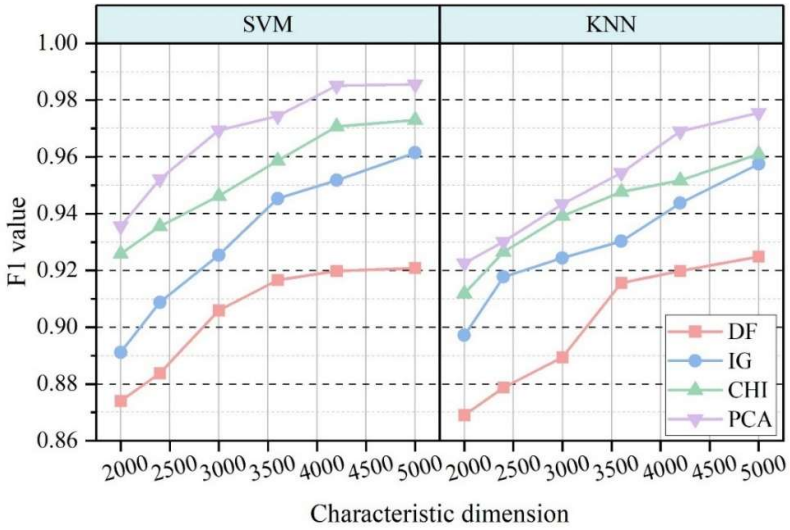

Fig. 2. Combine the classification results of two classifiers

In order to validate the classification effect of the proposed method on each category even further, the following chooses to show the F1 values of the PCA method as well as the DF, IG, and CHI methods on each category at the optimal feature dimension of 5000, respectively. In addition, according to the CPV $\geqslant$ 85% criterion, the feature dimension after dimensionality reduction was 4350, and the matrix was subsequently input into 2 classifiers respectively, with the parameter selection method remaining the same as above.

The results of comparing the classification effect of the natural imagery category of Tang poetry under each classifier are shown in Figure 3. From the data in the figure, it can be seen that all four methods reach the highest F1 value in the animal imagery category, because the animal imagery category in the corpus contains feature words with more category differentiation, which makes the classification effect better. Among the two classifiers, the PCA method has F1 values between 0.923 and 0.953, which are higher than the comparison method on all categories, with the highest improvement of 15.70%. This shows that the classification performance of the PCA method is also substantial on all categories.

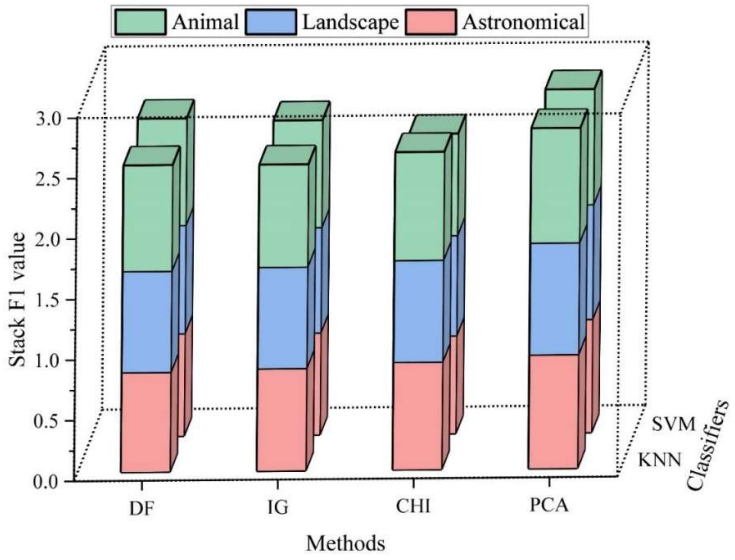

Fig. 3. The classification of the natural image category of tang poetry

In addition, the experiment also simultaneously compared the classification effects of different

methods for emotion recognition in different natural imagery. By mining the text of Tang poems, it is obtained that the emotions contained in the natural imagery therein include six kinds of emotions: longing, indifference, longing, joy, grief, and neutrality. The average F1 value of emotion recognition of natural imagery in Tang poetry under each classifier is shown in Figure 4. From the data in the figure, it can be seen that this paper's method exhibits the highest average F1 value for emotion categorization in different natural imagery on both classifiers. Taking the KNN classifier as an example, the sentiment classification performance of this paper's method on astronomical imagery, landscape imagery and animal imagery is 5.67% to 11.02% higher than the comparison method.

The above experiments demonstrate that the comprehensive performance of the principal component analysis method is all higher than the comparison method, and the effect is basically the same on both classifiers, so the SVM classifier will be used in the following study.

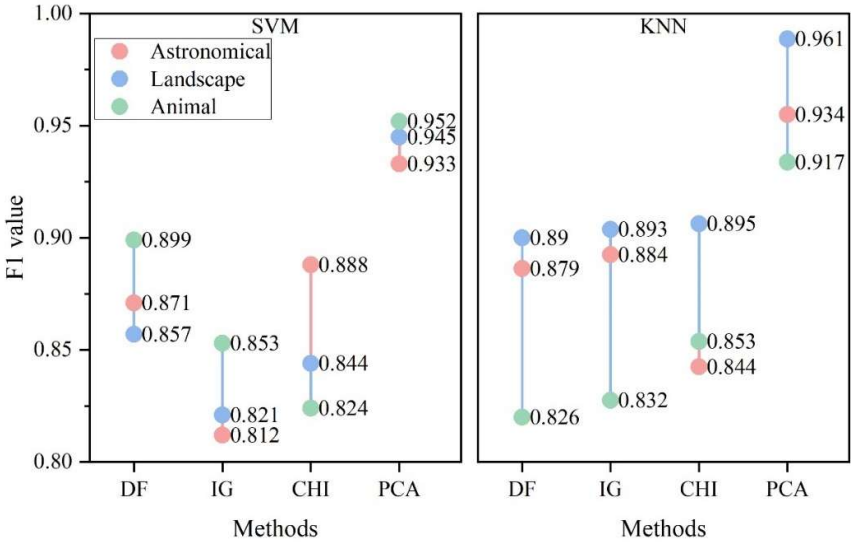

Fig. 4. The average f1 value of tang poetry's natural image

3.2. Analysis of Nature Imagery Groups in Tang Poetry

In order to facilitate the clustering analysis of Tang poetry texts and ensure the accuracy and reliability of subsequent experiments, while being able to express the natural imagery of the text. Based on the database of natural imagery of Tang poems established above, the experimental efficiency is improved by dimensionality reduction through the principal component model to obtain multiple key elements of natural imagery. With the semantic error allowed and without affecting the expression measure, based on the preliminary screening of 100 Tang poetry samples, SPSS mathematical statistics software was used to analyze the natural imagery clusters of Tang poetry samples.

The clustering results of the natural imagery clusters of Tang poetry texts are shown in Figure 5, and Table 1 describes the composition of each cluster class. According to the clustering results, the final 100 Tang poetry samples were categorized into 3 groups. Namely, astronomical imagery (RED), landscape imagery (BLUE), and animal imagery (GREEN), accounting for 38%, 53%, and 9%, respectively.

1) Astronomical Imagery. Astronomical imagery refers to symbolic natural things such as stars, moon, sun, etc. Ancient people also include some geomantic phenomena in astronomical imagery, such as wind, frost, rain, snow, thunder, fog, rainbow and so on. In ancient times, people's understanding of astronomy was very limited, for the existence of the sun, which has a very strong light, produces the feeling of more awe, while the temperature and light are softer the moon will be more generally loved and desired, the moon gives people a feeling of gentleness, good understanding, so Wang Wei would choose “moon” imagery That's why Wang Wei chose the imagery of the moon and wrote the poem

"The deep woods are unknown to the people, but the moon comes to illuminate them", a piece of his mind that has been quietly penetrated by the moon.

2) Landscape Imagery. All kinds of flowers, grass, trees and mountains are also common scenery in ancient Tang poems, they are not only the beautiful scenery in nature, but also constitute the unique flavor of the world of ancient Tang poems. For example, the pride of the plum, the elegance of the orchid, the integrity of the bamboo, the indifference of the chrysanthemum, and the purity of the lotus are all cultural values given to them because of their unique growth characteristics, and these landscape images with their own characteristics are the representatives of the natural imagery that poets can find in the mundane world to send their aspirations.

3) Animal Imagery. There are a lot of animal images in ancient Tang poems, some of these animal images are real in nature, and some of them are images of mythical beasts, which appear in literary works as an artistic image, often with a strong metaphorical and parasitic meaning. For example, in the poet's imagery world, geese imagery is often combined with other imagery to become the best thing to express the sadness of traveling.

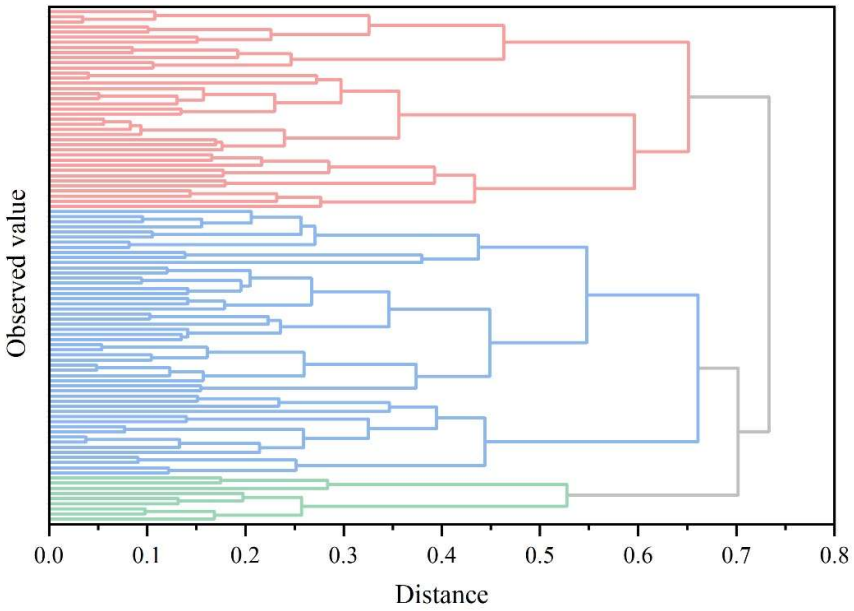

Fig. 5. The results of the group clustering of the text nature of tang poetry

Table 1. Composition of various clusters

Clusters	Image composition	Proportion
Astronomical image	Wind, frost, rain, snow, thunder, fog, rainbow, etc	38%
Landscape image	River water, peach blossom, bamboo, mountain, pear flower, pipa, etc	53%
Animal imagery	Cicada, bird, warbler, goose, crane, fish, horse, etc	9%

3.3. Nature Imagery Center Word Collocation Statistics

Based on the texts of Tang poems, this study aims to examine the discourse construction about nature imagery in Tang poetry texts using a combination of terminological principal component analysis and textual analysis.

This section investigates and counts the high-frequency nouns and word frequencies about nature imagery in the 100 Tang poems above. The statistical results of the top five high-frequency nouns and word frequencies of different nature imagery are shown in Figure 6.

From the figure, we can find that among the three types of natural imagery, astronomical imagery has the highest frequency, with a total frequency of 110 times for five types of high-frequency nouns, among which “wind” imagery appears 30 times, accounting for 27.27%. The imagery of “day” appears 15 times, accounting for 13.64%. The imagery of “moon” appears 17 times, accounting for 15.45%. Plant imagery appeared 104 times, of which “flower” imagery appeared 29 times, accounting for 27.88%. “Wood” imagery appears 24 times, or 23.08%. Animal imagery appears the least frequently, only 75 times, of which “mandarin ducks” and “swallows” appear 25 times and 19 times respectively, accounting for 33.33% and 25.33% respectively.

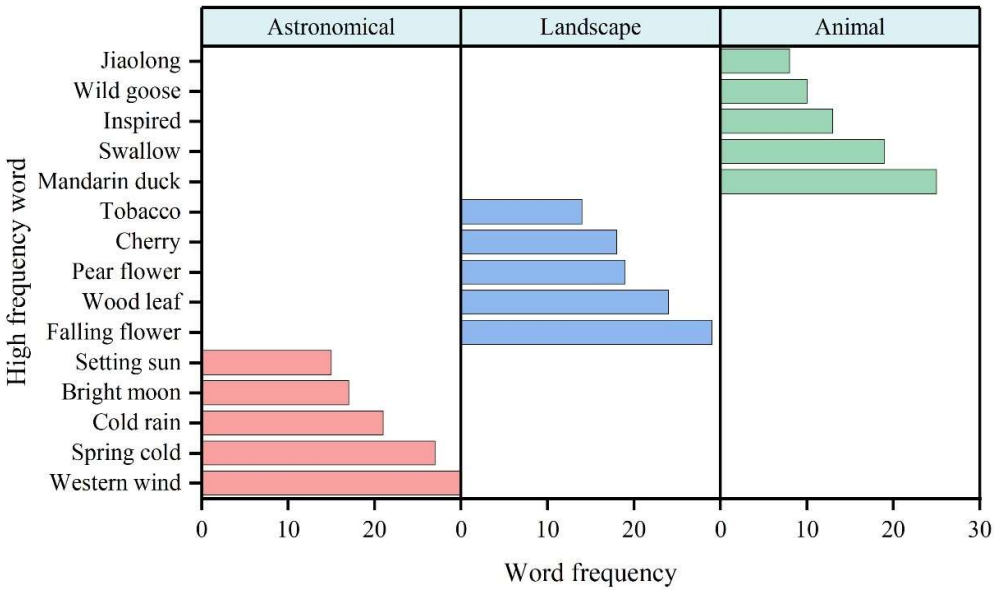

Fig. 6. Natural image high frequency noun 5 and frequency count statistics

Focusing on astronomical nature imagery discourse, the discourse field surrounding its occurrence was observed through collocation networks to measure the high-frequency collocations of this type of nature imagery discourse and the interrelationships between these words and nature imagery discourse. Figure 7 shows the co-occurrence analysis of the three categories of nature imagery theme words. The figure contains 27 natural imagery theme words, 1-27 representing, in order: blank, sad, a piece, past, dust, cold rain, life, old time, will rest, season, bright moon, this life, heavy sun, that year, full of land, west wind, earth, sunset, spring cold, like a dream, still, the end of the world, this day, countless, the east wind, this love, pear blossom.

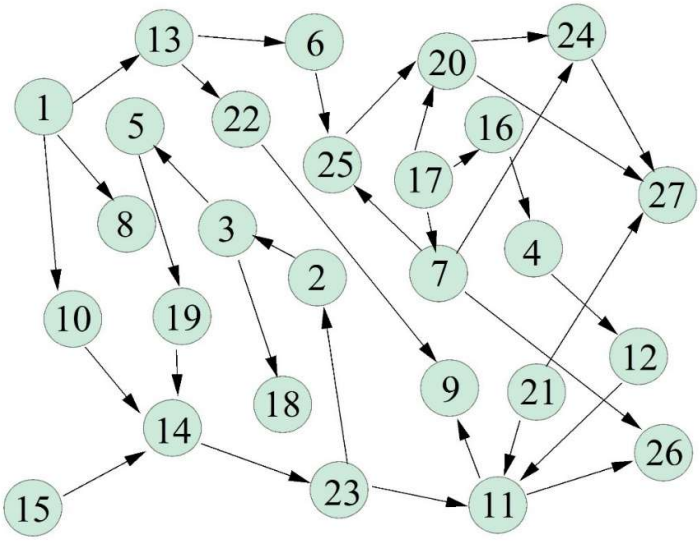

Fig. 7. The three types of natural images are analyzed

From the perspective of imagery, Table 2 shows the co-occurrence of words with collocation word frequency greater than 15. The purpose of analyzing the collocation word frequency of imagery word clusters and imagery center words is to analyze the references and connotations of the imagery as a whole. Through the above statistical results, it is found that: it is easier to form word clusters with the bright moon, the west wind, and the slanting sun as the center words, and the connotations are more fixed and clear, and it is easier to lock in the author's references to the imagery through the word clusters of the imagery. The words “old times - bright moon” and “looking back - west wind” point to longing, and are mostly used in remembrance of one's deceased wife. The combination of “Thoughts - Slanting Sun” and “The End of the World - West Wind”, whether imitating the tone of a girl in a bosom friend or describing the hardships of a military career, all point to the difficulty of feeling sadness. From the analysis of the four word clusters, the two clusters with the west wind as the core point to the expression of emotion, but with the different matching words, the emotional object of the word has changed. Therefore, a single image can only roughly outline the direction of emotion, and the specific emotional object needs to be analyzed in combination with the matching words and even the whole text. Although the images represented by east wind, dusk, and sunset appear with high frequency, the collocations are too loose and do not form a fixed cluster of image words, which makes it more difficult to grasp the meaning of the images, so it is necessary to go back to the context of the image words to find their definite meanings.

Table 2. The frequency of the combination of words is greater than 15

Central word	Collocation word frequency	Central word	Collocation word frequency
East wind	47	Western wind	44
blow	41	dusk	37
Setting sun	35	angling	34
Bright moon	30	swallow	25
Pear flower	23	Falling flower	21
Mandarin duck	17	Spring cold	15

3.4. Emotional Mining in Imagery

Emotional analysis of natural imagery is the process of analyzing and summarizing texts with subjective emotional colors. Tang Dynasty poems are very much in the style of the poet's personal

creation, containing personal emotional color and life experience. Through the text mining of Tang poems, the poet's emotional color and emotional tendency can be observed more intuitively. Text mining of emotional discourse in the dataset is carried out through the improved principal component model.

The results of the analysis of the emotional expression of natural imagery in the homemade Tang poetry dataset are obtained as shown in Figure 8. Statistical analysis revealed that for different natural imagery, poets expressed different kinds of emotions. The study investigated six kinds of emotions including longing, indifference, longing, joy, grief, and neutrality. Among them, neutral emotions were expressed in the smallest proportion (5.17% to 7.43%) of each natural imagery, indicating that poets would borrow natural imagery to depict their own poetic situations as a way to express the poet's emotions at this time. For example, Wang Wei's "The moon shines between the pines and the clear spring flows over the stones" shows the poet's emotion of transcending the world and returning to nature. Even for the same natural imagery theme words, the emotions that the poets develop can vary greatly depending on the circumstances of their poems and the way they express themselves. For example, "sunset" can express different emotions in different poems, such as Li Shangyin's "Ascending to Leyouyuan", which expresses the sadness of not being able to keep the beautiful things, while in other poems, it may express the affection for the travelers. Through principal component model analysis, a deeper understanding of how natural imagery in Tang poems is integrated with the poet's emotions, themes and styles can be achieved.

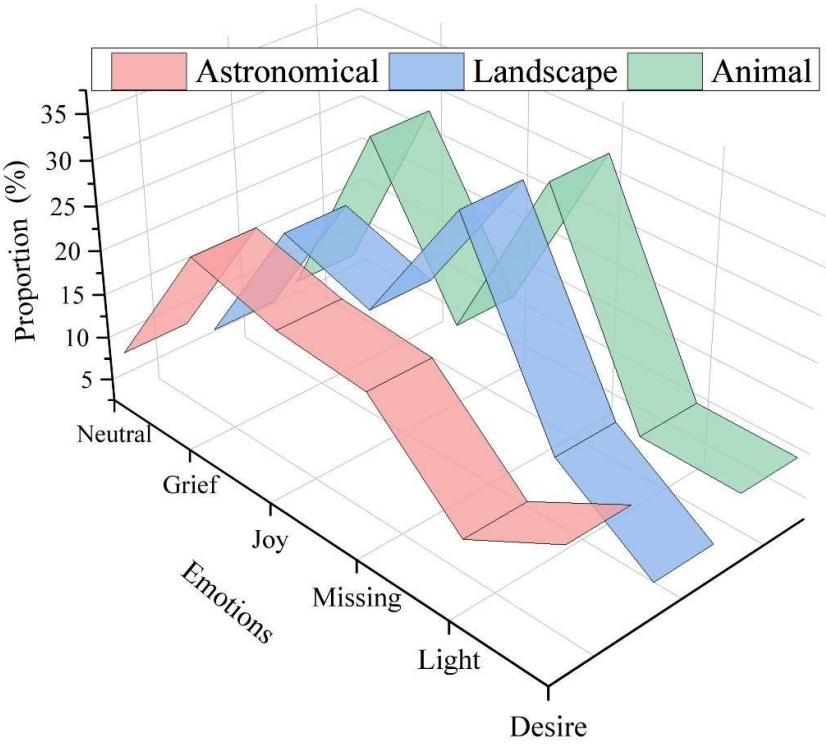

Fig. 8. Natural image emotion expression analysis results

4. Conclusion

Based on All Tang Poems, this paper constructs a natural imagery analysis dataset through text information extraction as well as data processing. In addition, an improved principal component model is proposed to efficiently reduce the dimensionality of Tang poetry text data by combining the algorithms of index homogenization and homogenization, aiming to improve the performance of classification of natural imagery and recognition of emotional expression in Tang poetry. On this basis, the k-means algorithm is used to classify the natural imagery clusters of Tang poems in the dataset,

and count the high-frequency nouns of natural imagery among them, and analyze the co-occurrence of various kinds of natural imagery theme words. The experimental conclusions are as follows: the principal component model combined with SVM and KNN classifiers has a better performance in classifying the natural imagery of Tang poems and their emotional expressions, and the F1 value is more than 0.9. After dimensionality reduction of the principal component model, the clustering algorithm classifies the natural imagery clusters of Tang poems into three categories, which are astronomical imagery, landscape imagery, and animal imagery, respectively. Among the three kinds of nature imagery, the astronomical imagery noun has the highest frequency of 110 occurrences. Moreover, the natural words of "Old times - bright moon", "Thinking - slanting sun", "Looking back - west wind ", "the end of the world - west wind" can better clarify the imagery expressed by the author. Combined with the principal component model analysis, the neutral emotion expression in the Tang poetry text accounts for the lowest percentage. To sum up, this study effectively utilizes information technology to mine valuable information from the massive data of this poetic text.

References

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

[2] Shelley, P. B. (2022). A defence of poetry and other essays. DigiCat.

[3] Wang, Q. (2017). The Expressive Forms of Natural Imagery in Chinese Poetry. Advances in Literary Study, 5(01), 17.

[4] Partner, J. (2018). Poetry and Vision in Early Modern England. London: Palgrave Macmillan.

[5] Liu, J. (2024, December). Research on the application of natural language processing in the analysis of ancient poems and texts. In AIP Conference Proceedings (Vol. 3194, No. 1). AIP Publishing.

[6] Al Matarneh, M. A., & Abuhammam, E. A. (2021). The images of nature in Wordsworth's and Al-Bohtory's poetry: A comparative study of selected poems. Theory and Practice in Language Studies, 11(11), 1384-1395.

[7] Clay, J. S., & Vergados, A. (2021). Teaching Through Images: Imagery in Greco-Roman Didactic Poetry (Vol. 450). Brill.

[8] Goldman, J. (2017). Modernism, 1910-1945: Image to Apocalypse. Bloomsbury Publishing.

[9] Al Bataineh, K. B., & Abuhammam, E. A. (2024). The Treatment of Nature in the Classical and Modern Arabic Poetry: A Stylistic Study. Kurdish Studies, 12(2), 2082-2092.

[10] Das, S. (2018). India, Empire, and First World War Culture: Writings, Images, and Songs. Cambridge University Press.

[11] Jiao, Y. (2022). Research on the Artistic Conception of Multimedia-Assisted Ancient Poetry Based on AI Technology. Scientific Programming, 2022(1), 8538503.

[12] James, S. (2023). Spinoza and the Poetic Imagination. Australasian Philosophical Review, 7(1), 9-27.

[13] Badizadeh, M. S., Mehrabani, J., & Ghotbi, L. K. (2021). Study and comparison of image developments in Persian poetry (from classical to contemporary). Propósitos y representaciones, 9(2), 97.

[14] Brooks, C. (2018). Modern poetry and the tradition. UNC Press Books.

[15] Aloraini, A. A. (2023). Depiction of European nature by an Egyptian poet: An aesthetic analysis of Ali Mahmoud Taha's poetry. Journal of Namibian Studies: History Politics Culture, 33, 2289-2315.

[16] Owen, S. (2020). Readings in Chinese literary thought (Vol. 30). BRILL.

[17] Krieger, M. (2019). Ekphrasis: the illusion of the natural sign. JHU Press.

[18] Charters, S. (2019). The poetry of the blues. Courier Dover Publications.

[19] Macpherson, J. (2022). Fragments of ancient poetry. DigiCat.

[20] Welsh, A. (2019). Roots of Lyric: Primitive Poetry and Modern Poetics. Princeton University Press.

[21] Damon, P. (2021). Modes of analogy in ancient and medieval verse (Vol. 15). University of California Press.

[22] Frosh, P. (2018). The poetics of digital media. John Wiley & Sons.

[23] Zarnigor, D. (2021). Phytonymic comparison as a means to create image in poetry. ACADEMICIA: An International Multidisciplinary Research Journal, 11(9), 693-698.

[24] Griffiths, M. (2017). The new poetics of climate change: Modernist aesthetics for a warming world. Bloomsbury Publishing.

[25] Mingchao Qi,JunQiang Zhao & Yan Feng. (2024). An optimized public opinion communication system in social media networks based on K-means cluster analysis. Heliyon(24),e40033-e40033.

[26] Katrina M Morgan,Erin V Feeney,Philip C Spinella,Barbara A Gaines & Christine M Leeper. (2025). Patterns of trauma-induced coagulopathy in injured children: A principal component analysis investigating endothelial, coagulation, and platelet biomarkers. The journal of trauma and acute care surgery(1),36-41.

[27] Yongqiang Han & Ruirui Liang. (2025). A high-altitude geomagnetic matching area selection approach based on geomagnetic information entropy and geomagnetic direction entropy. Measurement and Control(1),119-127.

[28] Gil ho Shin & Hyun Yang. (2025). Vessel trajectory prediction in harbors: A deep learning approach with maritime-based data preprocessing and Berthing Side Integration. Ocean Engineering119908-119908.