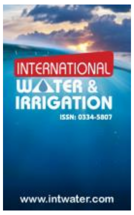


International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

Applying Text Mining Technology in Mining Thematic Trends in Chinese Literature

Yanfang Yan[1], Tao Liu[2*]

[1]Foreign Languages Department, North China Institute of Aerospace Engineering, Langfang, Hebei, 065000, China

[2*]Publicity Department, North China Institute of Aerospace Engineering, Langfang, Hebei, 065000, China

[2]*Corresponding author e-mail: liut0430@163.com

Abstract: Combining text mining technology with mining themes has emerged as a valuable tool for analyzing Chinese literature. This work presents a novel technique to quantify the features of text mining by merging mining algorithms. Determining the significance and language of ancient Chinese literature becomes less subjective as a result. This work demonstrates that the model can classify text mining words in Chinese literature accurately, up to the maximum degree of precision. The study proposes the use of text mining to analyze thematic trends in Chinese literature, employing a model known as the Enhanced Hierarchical Based Gradient Boost Algorithm (TEHGBA). For every Chinese literature outcome, the new TEHGBA model receives higher marks. According to the study's findings, the suggested model produced outcomes with 98.6% accuracy, 95.7% precision, and 90% recall. The study comes to the conclusion that the suggested model aids in the analysis of Chinese literary works and yields excellent recall, accuracy, and precision outcomes.

Keywords: Text Mining, TEHGBA, Chinese literature, thematic trends

1. Introduction

Text mining methods are used to get useful information and insights from text data that is not structured or is only partially structured. Text mining techniques have become very important for researchers and professionals in many fields because of the huge amount of digital material that needs to be analysed. It is possible to find patterns, trends, and connections in text data that would be hard or impossible to find by hand using mining method [1]. Text mining is the process of turning many unstructured text data into organized data so that you can get the information or knowledge you need from it. Chinese Literature is a field that colleges and universities offer to help students learn the Chinese language and literature [2]. By studying Chinese Literature, students can learn the basics of the field as well as news, history, art, philosophy, and other topics. They can also read classic works, do scientific research, do practical work, and so on. They can also learn about the latest achievements and future plans for colleges and universities. Traditional Chinese culture is an important part of Chinese literature, so the course materials must be based on and promote traditional Chinese culture [3]. In this way, Chinese literature becomes a propaganda tool through its teaching and indirectly spreads traditional culture. Each text and type of article should help learn what and how to gain, based on how the teacher sees the value and power of the language materials [4]. There are now a lot of Chinese language and literature teachers who are paying attention to how Chinese literature can help students learn the language, improve their skills, and learn new methods. Textbooks on Chinese literature have a lot of different types of texts and styles. Each style of text has its own features, like how the chapters are organized, the language used, and how it is expressed [5]. This study determines the thematic trends in the Chinese literature using the text mining technology.

Contribution

The aim of this study is to

- List out the different method using text mining to find research trends.

- Proposed a model for analyzing the data about Chinese literary works.
- Focus on what these methods can and can't do in terms of accuracy, speed, scalability, and generalizability.

2. Literature Review

In Chinese literature, there are also different reading styles and methods that are useful for teaching the language. Therefore, when looking into the teaching value of materials, it's important to pay some attention to the genre's traits and find out how different genres can be used to teach in different ways. Students can learn how to read by paying moderate attention to the genre's features and exploring the "class" of the genre's teaching value [6]. This will help them understand the language features of each "class" of the genre and give them access to the reading, thinking methods, and rules of that "class." The skill to grow over time can also be gained by teaching reading [7]. Large amounts of text data from many different sources, like news stories, scientific journals, and social media sites, are becoming easier to find. This has given investigators new ways to learn about new trends and topics in their field. However, analysing this kind of data by hand takes a lot of time and is prone to mistakes, biases, and limits [8]. Text mining methods can get around these problems by automatically pulling out useful data from text data and letting you analyse it in a way that is objective and based on data. There are a lot of different themes and emotional undercurrents in Chinese literature. This model is one of the most cutting-edge ways to understand these things [9]. This review aims to show how computational tools can be used to find hidden patterns, inter textual connections, and cultural influences in Chinese literary traditions by combining the newest research finds and methods from both computer science and literary studies [10]. In this paper, topic modelling was used to find research trends in the area of healthcare training [11]. They came up with eight main topics, such as online learning, multidisciplinary instruction, simulation-based learning and testing, and more. The study used clustering of texts to find trends in research in computer science. Six groups of study were found, including ML, NLP, data mining, and more [12]. The study used bibliometric evaluation and mining of texts to find research trends in tourism [13]. The most popular themes were found to be marketing destinations, the behaviour of tourists, and environmentally friendly tourism [14]. The study used both clusters and co-word analyses to find research themes in the area of big data analytics. They came up with six themes, such as data mining, cloud computing, and the internet of things [15]. The researchers used subject modeling and analysis of co-citations to find trends in the area of disaster management. They found four main research themes: preparedness for disasters, disaster reaction, and tragedy risk assessment [16].

3. Methodology

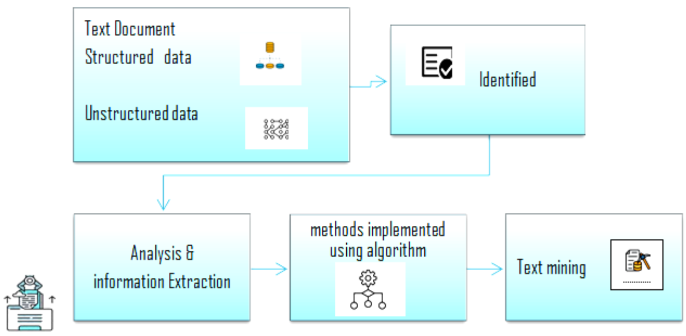


Figure1: Proposed Method

Figure 1 represents the proposed model of the study. This study centralized on the method is text mining. Structured and unstructured data are found in the text document. We need to identify the structured data in the text document. Then, extraction process will be taken place using the methods called TEHGBA.

International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

Text mining techniques

Using analytical tools to find and pull out possibly useful information from text that people are interested in is called text mining technology [17]. Natural language processing, data mining, ML, statistical analysis of data, probability theory and statistics with mathematical, linear algebra and analytical geometry, and even graph theory are all parts of this field. In order to do text mining, you need to gather text data, preprocess text, represent text, model text mining, and then test and use the model. Three of them are feature extraction, Chinese text segmentation, and Chinese text de-duplication. These are the three most important steps in getting a Chinese book ready. Text mining mostly talks about the theory behind semantic analysis, text representation, semantic networks, and topic modelling. This is because separating Chinese words is such an important part of text mining. In this study, four types of methods are analysed for mining thematic trends in the Chinese literature. The methods are BERT, LDA, TF-IDF and proposed method TEHGBA. Figure 2 represents the text mining models used in this study.

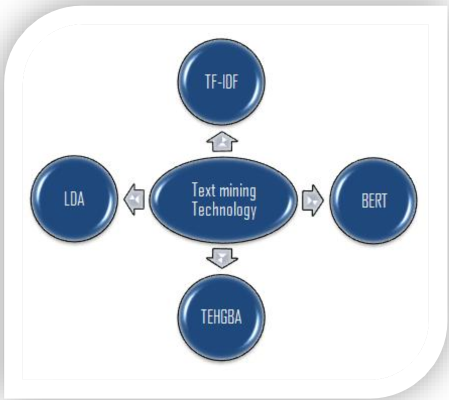


Figure 2: Text Mining Technology Model

3.1 TF-IDF algorithm

TF-IDF method has been improved. When words show a lot in a document, this is called TF. There is a big difference between these frequencies because different papers are different lengths. So, they can be compared in this method. Here is the method for how to find common TFs:

$$tf_{id} = \frac{m}{M} \tag{1}$$

Where equation (1) shows that m the total count of instance t appears in d . M is the overall total of words in d.

The opposite of frequency documentation IDF shows how important a word is; the common method is that show up less or more often in different types of documents. If a main feature shows up frequently in one type of record but not at all in the others, that means it is very good at telling the difference between the types and should be given more weight. When you need to figure out word frequency, you start by assuming that every word is important. At this point, a coefficient is needed to lower the weights of those words that are used a lot to balance their frequency. Here is the equation (2) for finding the shared IDF:

$$\frac{idf}{dt} = \log\left(\frac{N}{nt} + 0.01\right) \tag{2}$$

If N is the count of texts in the collection as a whole and T refers the count of word that contain the word highlight in t with +0.01, which keeps Tbperform from 0 to tiny collections.

In general, TFIDF method is according to the bunch-of-text, which says that text can be mentioned by a group of records that are,utilized it. The TFIDF theory says that a text should appear a lot in one record but not at all in others if it is important for that document. The first explanation is linked to TF, while the second explanation is linked to IDF.

Length of a feature term can be a key indicator of how important that word is. When Chinese words are separated into their parts, the data results show that single-character words appear more often than multi-character words. Single-character words can only say a few things, but different words of character can says lot of things that are greater significance.

Also, using the IDF formula, if there are m texts in one category of the corpus that contain feature word t and n texts in the other categories, then there are d papers in the corpus that contain feature word t. When m goes up, so does d and value of formula is less massive that was the problem with the standard TFIDF technique. On the other hand, this shows that the standard TF-IDF method is not good enough.

Based on the method, IDF strategy is improved, Then, the categories in setwith the C is represent corpus,$C = \{C_1, C_2, \dots \dots . C_x\}$ the group of words $C_m(C_m \in C)$in $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \dots \dots . \mathrm{d}_y\}$, and the set of main words is represented by $\mathrm{I} = \{\mathrm{i}_1, \mathrm{i}_2, \dots \dots . \mathrm{i}_z\}$, where x,y, z is the count of category, text with count and the count of main words, correspondingly,the equation(3) is illustrated as:

$$IDF' = \log\left(\frac{\sum_{d_1=1}^{T} m_{d_1} l_{d_1}}{\sum_{d_1=1}^{T} m_{d_1} l_{d_1} + \sum_{d_2=1}^{S} n_{d_2} l_{d_2}} \times n \sum_{d_3=1}^{K} l_{d_3}\right) \quad (3)$$

where $m_{d_1}$is the count of the words with times t appears in d_1 that belongs to group $C_i$and $d_1$is the text with lengthd_1. l_{d_2} is length with wordd_2, and T is the count of words that contain t in the category that belongs. $n_{d_2}$is the occurrence with frequency is t in words $d_2$is the other category besides category C_i. The count of texts which include t in the group to which it belongs is given by S. The count of texts that include t in the text is given by n. The length of the text is given by l_{d_3}, and K is the count of texts that include t out of all d.

Then the equation (4) combining the class with intra & inter is:

$$w(T_{ik}) = \frac{\mathrm{tf}(T_{ik}) * \mathrm{idf}(T_{ik}) *}{\sqrt{\sum_{k=1}^{n} (tf(T_{ik})) - [\mathrm{idf}(T_k)]}} *(1 - \mathrm{D_{ic}}) \quad (4)$$

3.2 BERT Analysis for the Chinese Literature

According to BERT analysis is Literature for Chinese, trained BERT models are used pull out situated embeddings with word and do tasks like mood analyse on Chinese text data [18]. The BERT's method, which has been taught on a large amount of text information, can understand words and their context. A transformer-based language model that has already been trained to pull out contextualized built in words and do jobs like sentiment analysis on Chinese text data. There are several important steps that go into making BERT Analysis for Chinese Literature.

The Chinese text is first broken down into sub words or characters, and then it is encoded into numbers that can be used by the BERT model. The compressed input sequence is shown by $X = [x1, x2, .., xn]$, where S is the length of the segment. Once these tokens are encoded go through a built-in model, they are turned into dense vector representations, which are shown by $E = [e1, e2, .., eT]$, where ei token is applied for vector xi. After that, the embeddings with input go several modes of Transformer encoders, which record information about the environment and the connections between tokens. Self-attention processes in each encoder layer calculate attention scores (l), which help the method to focus on the sequence of input.

International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

At first, the Chinese text is tokenized, which creates encoding tokens that are shown as $X = [x1, x2, .., xn]$, where n represents the length of the series. The encoded tokens are then put through BERT embeddings, which gives us $EBERT = [e1, e2, .., en]$, where ei is the BERT embedding for token xi. At the same time, lexicon-based sentiment analysis is used to give each token in the raw text a sentiment score $SLexicon = [s1, s2, ..., sn]$. When these emotion scores are added to the BERT embeddings, they create enriched representations that are shown as $\boldsymbol{E}Concat = [\boldsymbol{e}1 \oplus s1, \boldsymbol{e}2 \oplus s2, \ldots, \boldsymbol{e}n \oplus sn]$, where $\oplus$ means "join." After that, these representations are sent to LSTM layers, which makes it easier to find long-range relationships and patterns in the data that happen in a certain order.

3.3 Latent Dirichlet Allocation (LDA)

LDA is model that generates groups with datasets be clarified by groups that haven't been observed yet. These groups are the parts of the data alike. NLP and statistical ML have both been changed a lot by LDA. It has also famous probability-based text modelling method in ML very quickly [19]. As a result, each paper is seen as a mix of topics that apply to the whole corpus. A topic is a group of words that belong to the same language. These topics are made up from the papers that were collected. For instance, the words "football" and "hockey" are very likely to come up in the sports topic, while the words "data" and "network" are very likely to come up in the computer topic. Next, every word is assumed to originate from one of the themes in a set of papers that have an average probability across them. We can determine which themes are most frequently discussed and how much each issue is discussed in a document using this document probability distribution over each topic.

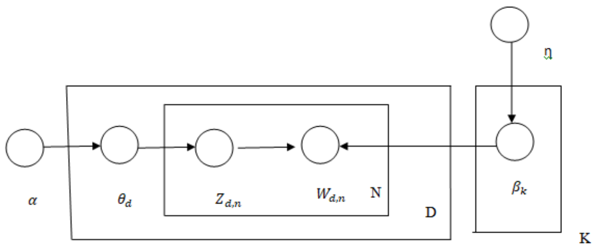


Figure 3: LDA Model

As the figure3 demonstrate the LDA the following notation. α and η are the amount specifics and topic specifics. The topics are denoted as $\beta_{1:K}$, where βk represents a span of the language. The topic allocation for topic k in document d is denoted by θ_{d}, k, which represents the topic amount for the dth document. Z_d, whereZ_d,n is the topic assignment for the n th word in document d, is the theme assignment for the d th document. In conclusion, the text that has been seen for document d is w_d, where w_d,n represents the nth writing in document d that is derived from the limited phrase. The LDA generative process is shown below, which shows how the hidden and visible variables are spread out when this equation (5) is used:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n})) \tag{5}$$

3.4 Data Preprocessing Using Normalization

The aims of data preprocessing are to make sure that only important data is kept, that each data block has only one piece of data, and that any data that is duplicated has been taken out. During this process, the independent variables are always changed in the same way to make sure they stay within a certain range. While there are other normalization methods that can be used, we choose to use Min-Max text mining in this case. The mining expression (6) looks after that:

$$n = \left(\left(\frac{(a - a_{min})}{(a_{max} - a_{min})} \right) * (1 - 0) + 0 \right) \quad (6)$$

In mining, the max and min signs show the highest and lowest numbers of the data, and the range is from 0 to 1. Values that have been noticed or estimated for an attribute show the average and mode of a character. Minimum skewness, which used to be described as the difference in a dataset's spectrum, means that the data sources are correct. As a result, Equations (7), (8), and (9) show the average, the middle point, and the variation.

$$M = \frac{\sum_{i-1}^{n} y_1}{n} \quad (7)$$

$$M_d = \begin{cases} y\frac{n}{2}, & if\ n\ is\ positive \\ \frac{\left(y\left[\frac{n-1}{2}\right] + y\left[\frac{n+1}{2}\right]\right)}{2}, & if\ n\ is\ negative \end{cases} \quad (8)$$

$$S_k = \frac{n \sum_{i=1}^{n} (y_i - \overline{y})^3}{(n-1)(n-2)\sigma^3} \quad (9)$$

In a collection, 'n' stands for the total number of values it contains, and 'y' stands for those values. The average is shown by $\overline{y}$and the standard deviation is shown byσ. To get security back to normal, all the necessary information needs to be found and gathered. In a number of ways to find outliers, attributes in normalization are very important.

3.5 Using Principal Component Analysis (PCA) To Figure Out Features

PCA is a frequently used method for analysing large datasets with numerous variables and features that can be interpreted in several ways. The goal is to keep as much of the original data as possible while making it simpler to understand. This is done with characters. For the PCA method to work; the feature set is what makes the Eigen values of the correlation matrix. A small group of traits, some with high Eigen values, will be split apart so that more research can be done. You can ignore the rest of the traits. Because of this, the huge depth of the subset of traits is made much simpler. This is one way to show the association matrix.

$$B = \frac{1}{x} \sum_{n=1}^{x} \{G_n - \alpha\} \{G_n - \alpha\}^T \quad (10)$$

In equation (10), Where $G_n$is the design (n =1 to x) the set of objects is $x,$ and the characteristic vector is x. With T standing for the inversion of the matrix.

$$CV_i = \tau_i u_i \quad (11)$$

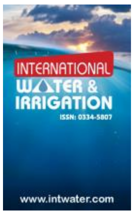
In Equation (11) (i = 1, 2, 3…n), n = number of features τ_i is the Eigen value and u_i is the Eigen vector.

We can figure out the discriminant separation $q_n^{x,y}$between subgroups x and y by knowing their standard deviations and mean values. Use the answer to equation (12) to find the characteristic value using PCA.

$$q_n^{x,y} = \frac{\left|(mean\ p_n^x) - mean(p_n^y)\right|^2}{[Sd\ (p_n^x)]^2 + \left[Sd(p_n^y)\right]^2} \quad (12)$$

Here, $p_n^x$and p_n^ystand for the nth attribute for transporting items under conditions x and y, respectively. The distance between the two groups of directions x and y for the nth capacity is given by $q_n^{x,y}$, where mean () and standard deviations () are the mean and standard deviation, respectively. The next equation (12) shows that the dispersion within a category is equal to the sum of its independent sectors. On the other hand, the dispersion between subcategories is equal to the squares of the differences between the means of those subcategories.

This is a 2D picture of a one-dimensional difference between two sets of data. There is data in equation (13). It is for the many (category both x and y) tests. It is clear what the difference is between the general mean and the test variances.



International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

$$T_n^{x,y} = \frac{|(Mean\ p_n^x) - Mean(p_n^y)|^2}{[Std\ (p_n^x)]^2 + [Std(p_n^y)]^2} \quad (13)$$

Where x and y are the respective weight limitations for the nth attribute and a component and n is the number that was used to separate those features. Use the info on the separation characteristics for each section to make an A-shaped matrix.

$$Y = [Y_1, Y_2, \dots \dots, Y_n] \quad (14)$$

As of this writing, there are n records, which means that it has r sets of data. The next step is to use equation 15 to find the association matrix S.

$$T = \frac{1}{m}\sum_{k=1}^{m}(Y_k - \bar{Y})(Y_K - \bar{Y})^S \quad (15)$$

The number Y is used here to show what I mean. The Eigen values of S are $[\delta_1, \delta_2, \dots\dots, \delta_n](\delta_1 \geq \delta_2 \geq \dots\dots \delta_2 \geq)$, the eigenvector S = [u1, u2,......, un] needs to be made. This Eigen vector gives business information a base that is not related to anything else. Things that have a higher value might be more useful. The fraction CD can be found using normalized methods, as shown in equation 16.

$$CD = \lambda k\left(\sum_{i=1}^{m} \delta\, j\right)^{-1} \quad (16)$$

Get rid of any Eigen values that does not add much to the attribute that was chosen. Using model C as a guide, the first d vectors are used to make the restoring vector Y. The automatic processes use a physical data set that puts more weight on the parts that are most reliable (as shown by the correlation matrix) than on the parts that are least reliable. This gives a way to cut down on the count of measurements, which makes the search for the best features go faster.

$$y = \widehat{\sum_{\iota=1}^{d} u_\iota^T} X u_\iota \quad (17)$$

3.6 Text Mining Based Enhanced Hierarchical Based Gradient Boost Algorithm (TEHGBA)

In this scenario, we have a training set with pairs like $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$; where xi is the instance's trait and $y_i \in \{0,1\}$is its feature. ML methods are used to score credit by making a variable $F(x_i)$ that minimizes the prediction error$L(y_i, F(x_i))$.

$$F *= arg \min_F \sum_{i=1}^{N} L(y_i, F(x_i)) \quad (18)$$

In Boosting gradient methods are used to solve Equation (18) through an extra optimal system:

$$F(x) = \sum_{t=1}^{T} f_t(x), \quad (19)$$

T is the total number of times this has been done. Using equation (19), we can see that the integration of $F(x_i)$is done more than once at each step. Iteration t makes the loss of all the aggregations made so far, $\{f_j\}_{j=1}^{t-1}$,, even better by usingft. TEHGBA uses an ML model to carry out each function f. This model can be thought of as a starting place for learners. This means that f may be represented as $f(\alpha; x), \alpha$where are the architectural characteristics of each ML and controls the attribute and dividing thresholds at every interior separating node in the ML approaches.

In this case, the loss functional is written as as $L(y_i, F_{t-1}(x_i) + f_t(x_i))$, and it is optimized at phase t th. Using Taylor series expansion to get a rough idea of the forecast error gives us the following:

$$L(y_i, F_{t-1}(x_i)) + g_i f_t(x_i) + \frac{1}{2} f_t(x_i)^2, \quad (20)$$

International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

The first derived of the loss functional, which is written as g_i, by solving:

$$g_i = \left[\frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{t-1}(x_i)} \tag{21}$$

So, the answer to equation (21) can be rewritten as the best answer:

$$f_t^* = arg\min_{ft} \sum_{i=1}^{N}(f_i,(x_i) - g_i)^2 \{y_i\}_{i=1}^{N} = -\left[\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{t-1}(x_i)}\right] \tag{22}$$

Equation (22) shows that the decreasing gradients of the function of error appear to be adapted to the goal of ft. This is the way we change the learning goals of each node while training them in TEHGBA:

4. Results And Discussion

4.1 TF-IDF

Document Category	Class Balanced Corpus (%)	Class unbalanced Corpus(%)
Finance	87.5	71.2
IT	85.9	68.6
Sports	93.45	72.4
Education	88.45	82.3

Table 1: Word frequency Statistics

After word segmentation is done on the text data, the long comment lines are broken up into words. The model in this paper's character frequency's statistics function can then be used to count the high-frequency words in the communication. The value of the character chance is the times with count that the word found in every text. We use the word count of times as a guide. In this work the TF-IDF algorithm used to find out how often words are used. The results are shown in table 1and Figure 4.

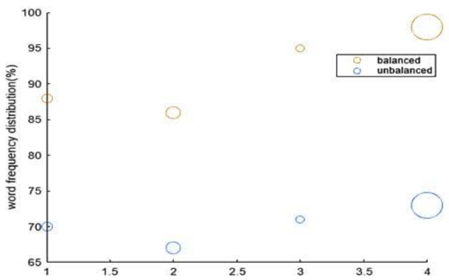


Figure 4: TF-IDF Word Segmentation Frequency

4.2 BERT Analysis

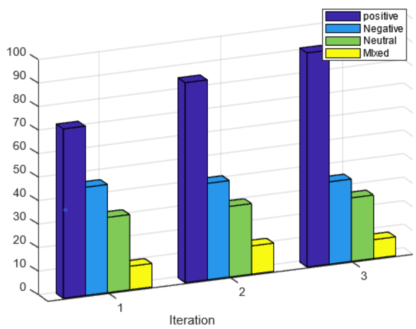


Figure 5: Classification With Iteration

The way of model work was tested over a lot of versions and found to be consistently accurate at classifying feelings. The model constantly performs well across five iterations, with accuracy values above 0.93. This shows that the model is good at correctly identifying feelings in Chinese literature texts. Also, the accuracy value is always between 0.94 to 0.95, which shows that the model can correctly identify positive, negative, and neutral emotions while reducing the count of false positives. According to the same research, recall values stay between 0.93 and 0.94, which means that the model can pick up a good number of true positives out of all real positives [14].

4.3 LDA

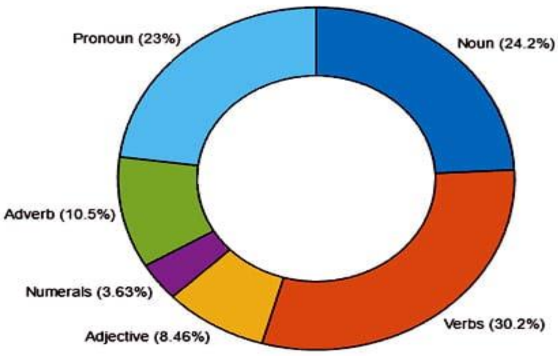


Figure 6: The Vocabulary Density in The Text

In figure 6, we can see bajin's Cold Nights and Rest Garden. The text of Cold Nights has a slightly higher vocabulary density than that of Rest Garden by 2.1 percentage points. The two works have the highest verb vocabulary density, at 27.31% and 28.25%, respectively. There are a lot of dialogues in both books that show how the people talk. The word "say" is used 1,213 times in the first book and 735 times in the second. Text mining technology can successfully look at the language and vocabulary of Chinese literary texts, make the emotional expressions in Chinese literary texts clearer, and help us figure out why the author wrote anything in the first place.

4.4 TEHGBA

The general word count table is shown in Table 2. It shows the total count of words and the count of times each word appeared in each of the four sections that were looked at. There were 5901 different words in these 72 stories. The first thing we can say is that these 72 stories had more unique words than unique characters. This is because the same character can be used in more than one word, so there are more words made up of characters than there are characters themselves. Another result that seems to make sense is that the articles had fewer words (21071) and more characters (38080), there were fewer words than characters in these 72 stories.

There is a third result that supports the idea that the lot of words you know is more important than the count of characters you know. Characters stand for morphemes first and foremost, not words. Even though there are a lot of words with only one character, most Mandarin words have at least two. The character frequency count showed that there were only 2,100 unique characters in the 72 stories.

Table 2: Analyse The Words in Unique Articles

	No. of words	No. of different words	No. of articles
國際	4360	1782	18
政治	5141	1990	18
社會	6325	1935	18
財經	5245	1931	18

International Water and Irrigation
Volume 44 Issue 2 (2025)
ISSN: 0334-5807
Pages 92 - 103

Total	21071	7638	72

While it is true that it is easier to guess the meaning of a new multicharacter word made up of characters we already know, it is still easier to guess the meaning of a word with characters we have never seen before. we are familiar with the characters for ice (冰) and box (箱), and then see the word 冰縱, which means refrigerator, you might be able to figure out what it means. While we might be confused about whether it's a freezer, a refrigerator, or some other type of food-cooling box, it should be pretty clear from the rest of the sentence what it means. That being said, this isn't always the case. If you look at the characters for electricity (黻) and look (視) and see the word 黻視, you might not be able to figure out that it means television even after reading what it was used for.

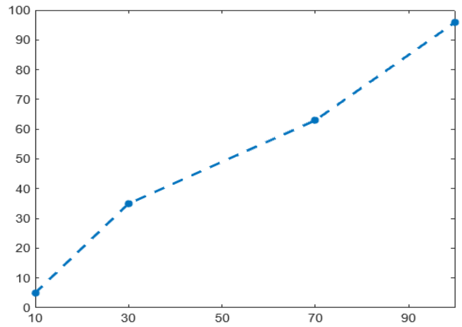


Figure 7: Character Frequency Distribution

4.5 Models Comparison

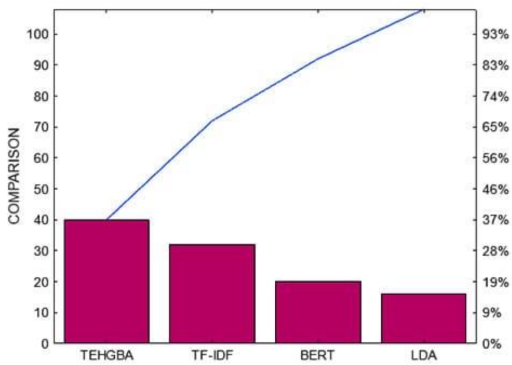


Figure 8: Models Comparison

When compared with all the models, TEHGBA shows high performance comparatively with TF-IDF [20] ,LDA [19] and BERT [18] the Chinese literature based on text mining technology.

4.6 Accuracy, Precision, Recall

Table 3: Performance Of Methods

Methods	Accuracy (%)	Precision (%)	Recall (%)
TF-IDF	72.5	70	63
BERT approach	67.8	66.2	65.5
LDA	54.9	52.3	51
TEHGBA[Proposed]	98.6	95.7	90

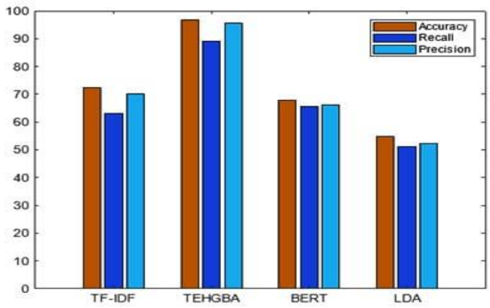


Figure 9: Accuracy, Precision, Recall

Figure 9 and Table 3 show the classification success metrics that were found using all four models over a number of different runs. Each row shows a different version, and the columns show accuracy, precision, recall, and, which show how well the model works at classifying things. When you look at the data, it's clear that the TEHGBA model always gets high accuracy scores, ranging from 90% to 100%. This shows that the model can correctly group text mining with literature in Chinese with a maximum level of precision. Also, precision values always go above 95, which show that the model can identify positive, neutral, and negative emotions predicted with minimum false positive. Furthermore, recall scores stay high, ranging from 90 to 95, showing that the model is good at catching a substantial portion of true positive instances out of all real positive instances. Strong performance of the model shows that it can read and understand complicated text data, giving us useful information about how people feel in Chinese writing. In addition, using BERT, LDA, TF-IDF, and TEHGBA together was especially helpful because it let the model pick up on both the meaning and contextual details of the texts, which led to more accurate text mining classifications. With these results, we can see how advanced text mining technology models can help us get deeper insights from textual data. This will help us understand how Chinese literature changes over time and could also help with research into markets, social networking analysis, and mining opinions.

5. Conclusion

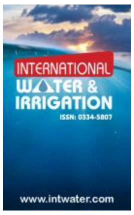
The study results show that text-mining technology can help people understand and analyze the complicated features of Chinese literature. This improves the way we study Chinese literature and gives us new ways to look at it and tools for using text mining in Chinese literature. This research is the first in-depth look at how to use text mining to keep up with the latest developments in design research. The result also shows that the methods that were created are universal and can be used to organize knowledge from many different study areas. Text mining methods used in this study could also help other researchers get a full picture of the knowledge about a certain subject that is hidden in a lot of Chinese literature. Therefore, the research presented in this paper offers new TEHGBA methods and new views for the text mining technology of Chinese literature. It also helps the field grow.

Reference

[1] Arshad, M., Khan, A., Ahmed, P. and Nadia Abbas Shah Next Generation Data Analytics: Text Mining in Library Practice and Research. Library Philosophy and Practice (ejournal), 2020,4768(1).

[2] Atandoh, P., Zhang, F., Adu-Gyamfi, D., Atandoh, P.H. and Raphael Elimeli Nuhoho, Integrated deep learning paradigm for document-based sentiment analysis. Journal of King Saud University. Computer and information sciences/Mağalaẗ ğam'aẗ al-malīk Saud : ùlm al-ḥasib wa al-ma'lumat, 2023, 35(7), pp.101578–101578. doi:https://doi.org/10.1016/j.jksuci.2023.101578.

[3] Cao, Y., Sun, Z., Li, L. and Mo, W. A Study of Sentiment Analysis Algorithms for Agricultural Product Reviews Based on Improved BERT Model. Symmetry, 2022, 14(8), p.1604. doi:https://doi.org/10.3390/sym14081604.



[4] Chen, M.-C. and Ho, P.H. Exploring technology opportunities and evolution of IoT-related logistics services with text mining. Complex & Intelligent Systems, 2021, 5(1). doi:https://doi.org/10.1007/s40747-021-00453-3.

[5] Chen, Y., Zhang, H., Liu, R., Ye, Z. and Lin, J. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 2019, 163(5), pp.1–13. doi:https://doi.org/10.1016/j.knosys.2018.08.011.

[6] Hu, F. and Trivedi, R.H. Mapping Hotel Brand Positioning and Competitive Landscapes by text-mining user-generated Content. International Journal of Hospitality Management, 2020, 84(5).

[7] Konstantinos Vavousis, Papadopoulos, M., Michalis Gerolimos and Xenakis, C. Text and Data Mining for the National Library of Greece in consideration of Internet Security and GDPR. Qualitative and Quantitative Methods in Libraries, 2020, 9(3), pp.441–460.

[8] Liang, B., Su, H., Gui, L., Cambria, E. and Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowledge-Based Systems, 2021, 6(3), p.107643. doi:https://doi.org/10.1016/j.knosys.2021.107643.

[9] Liu, H., Chen, X. and Liu, X. A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis. IEEE Access, 2022, 10(3), pp.32280–32289. doi:https://doi.org/10.1109/access.2022.3160172.

[10] Mishra, S., Choubey, S., Choubey, A., Yogeesh, N., Durga Prasad Rao, J. and William, P. Data Extraction Approach using Natural Language Processing for Sentiment Analysis. [online] IEEE Xplore. 2022. doi:https://doi.org/10.1109/ICACRS55517.2022.10029216.

[11] ONAN, A. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. Computer Applications in Engineering Education, 2020, 4(2). doi:https://doi.org/10.1002/cae.22253.

[12] Onan, A. and Korukoğlu, S. A feature selection model based on genetic rank aggregation for text sentiment classification. Journal of Information Science, 2016, 43(1), pp.25–38. doi:https://doi.org/10.1177/0165551515613226.

[13] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W. and Yu, S. A survey on deep learning for textual emotion analysis in social networks. Digital Communications and Networks, 2021, 5(6). doi:https://doi.org/10.1016/j.dcan.2021.10.003.

[14] Raeesi Vanani, I., Mahmoudi, L., Jalali, S.M.J. and Pho, K.-H. Using text mining algorithms in identifying emerging trends for recommender systems. Quality & Quantity, 2021, 4(5). doi:https://doi.org/10.1007/s11135-021-01177-9.

[15] Shen, X. Sentiment Analysis of Modern Chinese Literature Based on Deep Learning. Deleted Journal, 2024, 20(6s), pp.1565–1574. doi:https://doi.org/10.52783/jes.3075.

[16] Song, G., Wu, J. and Wang, S. Text Mining in Management Research: A Bibliometric Analysis. Security and Communication Networks, 2021, 2021(4), pp.1–15. doi:https://doi.org/10.1155/2021/2270276.

[17] Wu, Y. Research on Text Value and Linguistic Characteristics in Ancient Literature Based on Text Mining Technology. Applied mathematics and nonlinear sciences, 2024, 9(1). doi:https://doi.org/10.2478/amns-2024-0390.

[18] Zhang, T., Li, B. and Hua, N. Chinese cultural theme parks: text mining and sentiment analysis. Journal of Tourism and Cultural Change, 2021,1(2), pp.1–21. doi:https://doi.org/10.1080/14766825.2021.1876077.

[19] Zhang, W., Wang, H., Song, M. and Deng, S. A method of constructing a fine-grained sentiment lexicon for the humanities computing of classical chinese poetry. Neural computing & applications, 2022, 35(3), pp.2325–2346. doi:https://doi.org/10.1007/s00521-022-07690-8.

[20] Zidan, M., Elhenawy, I., Abas, A. and Othman, M. TEXTUAL EMOTION DETECTION APPROACHES: A SURVEY. Future Computing and Informatics Journal, 2022, 7(1), pp.32–58. doi:https://doi.org/10.54623/fue.fcij.7.1.3.