Hindawi
Mobile Information Systems
Volume 2022, Article ID 9382890, 13 pages
https://doi.org/10.1155/2022/9382890

*Research Article*

# English Text Analysis System Based on Genetic Algorithm

**Jian Yao and Jin Xu**

*School of Foreign Languages and Literature, Tianjin University, Tianjin 300350, China*

Correspondence should be addressed to Jin Xu; amandaxujin@tju.edu.cn

Received 2 June 2022; Accepted 24 June 2022; Published 11 July 2022

Academic Editor: Jiafu Su

Copyright © 2022 Jian Yao and Jin Xu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to accurately extract the useful information in English, this paper studies English text analysis combined with a genetic algorithm and establishes a text analysis system. In this method, a text tendency analysis algorithm based on a genetic algorithm language model is proposed, and a Doc2vec text feature representation algorithm integrating the LDA model is designed; the parallelization technology of text analysis algorithm is studied, and the parallelization model of the algorithm by using spark big data platform is designed; the process of English text tendency analysis is studied, and a Chinese text analysis system is designed and implemented based on big data platform, including corpus intake, corpus annotation, corpus storage, model training, model verification, and other modules. In order to verify the feasibility of this subject, the accuracy of the Doc2vec text feature representation algorithm of the fused LDA model designed in the prototype system is tested. The experimental results show that the fused text representation model has high recognition degree, and the AUC value of the ROC curve reaches 0.95. At the same time, this paper tests the text analysis-related algorithms involved in the system. The test results show that the parallel algorithm can greatly improve the efficiency of the system.

## 1. Introduction

Since the birth of the genetic algorithm in the 1970s, many institutions and researchers have conducted extensive and in-depth research on it, achieved many important research results, and rapidly extended its application fields to optimization, search, machine learning, and other aspects. It has gradually developed into a calculation model to solve optimization problems by simulating the natural evolution process [1]. Content-based text information filtering is an important part of machine learning. Genetic algorithm was first applied to machine learning to solve some simple learning problems. For example, the CS-1 system proposed by Holland and Reitman applies the genetic algorithm to solve maze problems for the first time while Goldberg applies the genetic algorithm to engineering control. These studies have produced genetic-based machine learning (GBML) [2]. Text analysis is about the representation of text and the selection of its product features. Text analysis is a key problem in text retrieval and archiving. It counts words extracted from text to represent information as shown in Figure 1. Text has much the same meaning as text. It refers to a data format that contains symbols or numbers. These templates are available in multiple languages such as text and graphics. Words are created by special people, and the content of a book should reflect people's work, thoughts, values, and interests. Thus, the reader's purpose and intent can be determined by identifying the content of the text and converting them from nonstandard texts into documents that computers can recognize and process, i.e., study the texts and develop their mathematical models to interpret and alter the texts. Through the calculation and operation of the model, the computer recognizes the text [3]. Since text is data-intensive, in order to mine useful data from multiple texts, the text must first be converted into process code. Most people use a vector space model to describe vector text, but if the product features come from word segmentation algorithms and word frequency statistics used to represent vector text of different lengths, the length of this vector will be large. This unfinished business not only brings huge overhead to the next task, making the whole process inefficient, but also increases the accuracy of the distribution and grouping algorithms, resulting in unsatisfactory results [4]. Therefore, we need to further refine the vector text by refining the

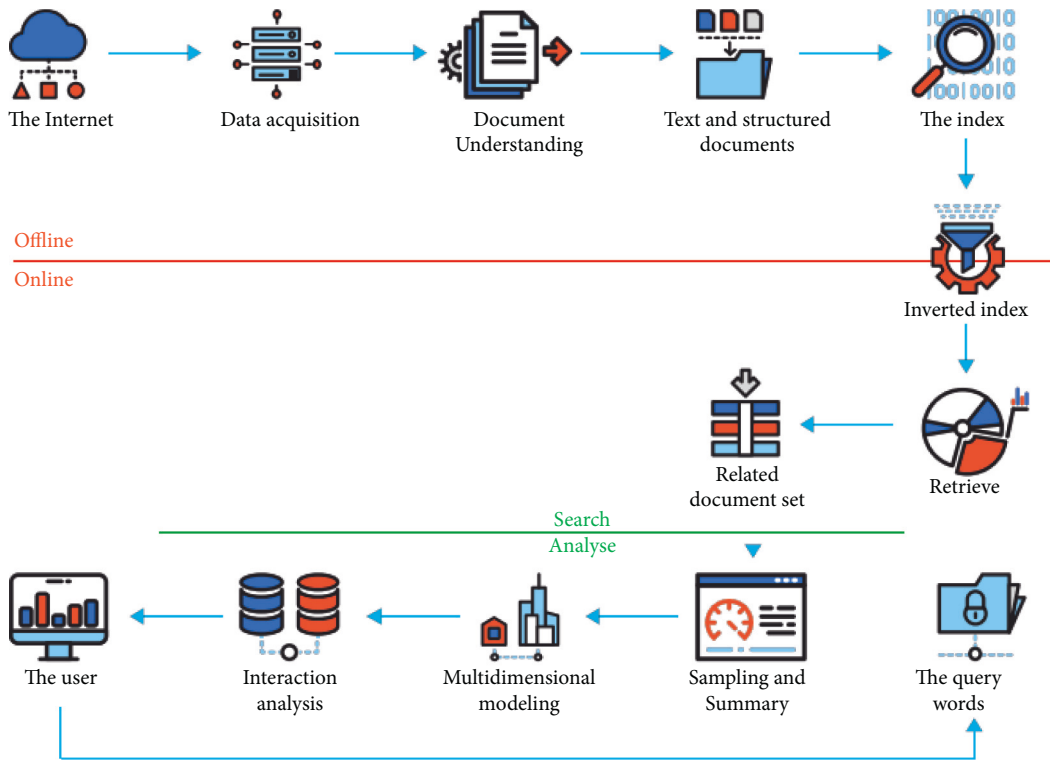


Figure 1: This analysis system.

content of the original text to find the most representative text for the text. To solve this problem, the best way is to reduce the dimensionality by selecting a feature.

## 2. Literature Review

Genetic algorithm is rarely applied to the research of text information generation, especially the application of the genetic algorithm to text information selection. In 2000, Wang and others first applied the genetic algorithm to feature selection [5]. Then, Liu and others introduce the text distribution region selection feature based on the genetic algorithm. Since then, many scientists have devised various improvements [6]. Bai et al. also discuss mutation algorithms and use them to select options [7]. According to Sheluhin et al., the distribution of text is the process of dividing the text into one or more subgroups according to the content. It is an expression. As a set of information training tools, find the relationship model between data features and data categories. These relational models are then used to determine categories of anonymous data [8]. Izrailova and Badaeva believe that in English text classification, the text set becomes a word set after word segmentation, and then, the feature set is obtained by removing the stop words and roughly reducing the dimension. However, the feature set is still a high-dimensional feature space, which is too large for all classification algorithms. Therefore, we are faced with finding an efficient feature extraction method to reduce the dimensionality of feature regions and improve the efficiency and accuracy of distribution [9]. Mufti and others said that the purpose of feature selection is to remove the features that cannot better represent effective information in the feature set so as to improve the classification accuracy and reduce the computational complexity [10]. Meng and others said that in text classification, generally speaking, when the text is expressed in vector form, there may be tens of thousands of feature items in the training text set. It is generally believed that any one of these features has its contribution to the realization of correct classification. However, these large number of features must also contain many interrelated features, which are redundant and can be removed [11]. Tominaga and others believe that too large feature space will make the evaluation of sample statistical characteristics more difficult, thus reducing the generalization ability of the classifier and causing the phenomenon of "over learning." Moreover, the processing of this high-dimensional vector has extremely high computational complexity, especially the so-called "dimension disaster" problem [12]. Therefore, Yue and Wang said that how to retain those features that play an important role in classification and remove redundant features in order to reduce the total number of features, that is, how to carry out dimension reduction, has become an increasingly important research field [13]. Shi described that the representation of text distribution as a process in data filtering, data recovery, archival, digital library, and e-mail distribution has wide application reliability [14].

## 3. Method

*3.1. Feature Selection Dimension Reduction.* Feature selection dimensionality reduction is classified according to the concept of algorithms and can be divided into three categories: filtering feature selection, encapsulation feature

selection, and embedding feature selection. The filtering feature selection algorithm is independent of the classification algorithm. It directly judges the advantages and disadvantages of text features according to the characteristics of text data and finally selects the excellent text features as the final feature subset. The encapsulated feature selection algorithm needs to preset the classification algorithm to obtain the classification model and indirectly evaluate the classification efficiency of the feature subset by detecting the final effect of the model on the test set. Embedded feature selection algorithm is used to automatically select features in the process of training classification model [15]. The commonly used filtering text feature dimensionality reduction algorithms are described in detail below.

*3.1.1. Term Frequency-Inverse Document Frequency (TF-IDF).* TF (term frequency) means word frequency in Chinese and IDF (inverted document frequency) means inverse text frequency index in Chinese. The theoretical focus of its algorithm is applied to text feature selection, which is to calculate the reciprocal product of the number of times a single text feature appears in the overall text set and the number of documents [16]. The calculation is shown in the following formula:

$$TF - \mathrm{IDF}_a = \frac{N_{a,A}}{N_A} \cdot \log \frac{Z}{Z_a + 1}, \tag{1}$$

where $N_{a,A}$ represents the number of times the text feature $a$ appears in document $A$, $N_A$ represents the total number of words in document $A$, $Z$ refers to the total number of documents in the corpus data, and $Z_a$ refers to the number of documents including text feature $a$.

When using the term frequency-inverse document frequency as the method of text feature selection, there are mainly the following two disadvantages: (1) when the data set is skewed, the calculation method of inverse document frequency will be affected by the imbalance of the number of documents, which is difficult to achieve our desired goal, and (2) the scoring standard only takes the contribution of text features to the whole as the weight, ignores the performance ability of text features in a single category, and the ability to distinguish between categories is weak [17].

*3.1.2. CHI Square Statistics.* After calculating the CHI squared value for letters and categories, the CHI squared value is calculated from large to small. The higher the value, the better the relationship. The CHI square value of text feature $a$ and category $P$ is calculated as shown in the following formula:

$$x^2(a, P) = \frac{N * (AD - BC)^2}{(A + C) * (B + D) * (A + B) * (C + D)}, \tag{2}$$

where $A$ is the number of documents with text feature $a$ and belonging to category $p$; $B$ is the number of documents with text feature $a$ but not belonging to category $P$, $C$ is the number of documents without text feature $a$ but belonging to category $P$, $D$ is the number of documents without text feature $a$ and not belonging to category $P$, and $N$ is the number of total documents [18]. For multiple distributions, count the squared CHI values for each format contained in the corpus data, and then, assign a mean or higher value based on the squared CHI value of text $a$. Studies have shown that the squared CHI statistic for the largest breast in multiple distributions is superior to the squared CHI statistic in terms of time and effect [19].

The CHI square statistical algorithm for text feature selection mainly has the following two disadvantages: (1) because the CHI square statistical algorithm based on interclass discrimination does not consider the competition between similar feature words, for example, it does not consider the interference of word frequency distribution between each type of feature words, which reduces the accuracy of its evaluation features and exaggerates the role of low-frequency words; (2) because there is the factor $(AD - BC)^2$ in the formula, if $AD < BC$ occurs in multiclassification, the characteristic words with poor classification effect will be wrongly given high score evaluation, which will interfere with the evaluation expressiveness of characteristic words.

*3.1.3. Mutual Information (MI).* Mutual information algorithm is a statistical algorithm that shows the correlation between two subjects. In text feature dimensionality reduction, it calculates the relationship between text features and corpus categories. It is used for text feature selection. It is usually used to judge the amount of information associated with text features and various categories as shown in the following formula:

$$\begin{aligned} MI(a, c_j) &= \log \frac{p(a, c_j)}{p(a) \cdot p(c_j)}, \\ &= \log \frac{p(a|c_j)}{p(a)}, \end{aligned} \tag{3}$$

where $p(a, c_j)$ represents the probability of the existence of text feature $a$ in Category $c_j$, $p(a)$ represents the probability of the existence of text feature $a$ in the total number of documents, $p(c_j)$ represents the probability of the existence $c_j$ of category in the total number of documents, and $p(a|c_j)$ represents the probability of the existence of text feature $a$ in category $c_j$. Let $\{c_1, c_2, c_3, \ldots, c_n\}$ represent the collection of categories in the document, then the mutual information calculation of text features in corpus data is shown in the following formula:

$$MI(a) = \sum_{i=1}^{n} P(c_i) MI(a, c_i). \tag{4}$$

Computational approaches to data sharing are always algorithms with the following disadvantages: (1) ignoring the influence of word frequency factor on features, the formula focuses on the selection of low-frequency words, resulting in the loss of some feature words with high word frequency and strong classification; (2) the feature may have a negative

value for the mutual information calculation value of a single category, indicating that the text feature does not exist or exists less in this category, which plays an important role in category judgment, and the value of the data cannot be affected in the computational model of the data sharing algorithm; (3) when the resources of various documents in the data set are unbalanced, the evaluation of text features is not accurate.

*3.1.4. Information Gain (IG).* The information gain algorithm takes the value brought by the feature to the whole as the evaluation standard and represents the amount of information brought by the feature to the system according to the difference between the amount of information when the system includes feature $a$ and does not include feature $a$ [20]. When calculating the information gain of a single text feature, calculate the difference between the direct line of the classification system when the text feature $a$ is included and the direct line of the classification system when the text feature $a$ is not included, which represents the information gain brought by the text feature $a$ to the classification system and the contribution value of the text feature $a$. There are two cases without feature $a$: the first case is that feature $a$ does not exist in the classification system, and the second case is that feature $a$ exists but $a$ has been fixed in the classification system. In the actual calculation process, we use the second method to calculate. At this time, the amount of information is also called "conditional Di," and the condition is that the feature $a$ has been fixed. In the Chinese text classification system, when feature $a$ is fixed, there are two situations: occurrence and nonoccurrence. We use $a$ to represent the occurrence of feature $a$ and $\overline{a}$ to represent the nonoccurrence of feature $a$. Let $\{c_1, c_2, c_3, \ldots, c_n\}$ represent the collection of categories in the document and $p(c_i)$ represent the distribution probability of various texts. The calculation is shown in the following formula:

$$\begin{aligned} IG(a) = &-\sum_{i=1}^{n} P(C_i)\log_2 P(C_i) \\ &+ P(a)\sum_{i=1}^{n} P(C_i|a)\log_2 P(C_i|a). \end{aligned} \tag{5}$$

The disadvantages of using the information gain method to evaluate the weight of text features are as follows: (1) paying too much attention to document frequency and weakening the attention to word frequency; (2) when the resources of various documents in the data set are unbalanced, the actual evaluation of text features will deviate from the expectation, resulting in inaccurate evaluation [21].

*3.2. Language Model and Text Representation Method.* Language model is used to model natural language. The traditional language model is a statistical language model, which is a probability distribution function representing language fragments. Its mathematical expression is as follows:

$$p(W) = p\left(w_1^T\right) = \Pi_{t=1}^{T} p(w_t|\text{Context}), \tag{6}$$

where $W = w_1^T = (w_1, w_2, \ldots, w_T)$ represents the language fragment composed of $T$ words $w_1, w_2, ..., w_T$ in order and $p(W)$ represents the probability of these words being combined. According to the Bayesian formula, the $p(W)$ chain can be decomposed into the following equation:

$$p\left(w_1^T\right) = p(w_1) \cdot p(w_2|w_1) \cdot p\left(w_3|w_1^2\right) \cdots p\left(w_T|w_1^{T-1}\right). \tag{7}$$

If the context of each word is uniformly recorded as context, the expression of equation (6) can be obtained. Different language models can be formed according to different context division strategies. The design usually includes the N-gram structure, n-pos structure, decision model, maximum entropy model, maximum entropy Markov model, and neural network language model. Different text modeling methods will have different effects. At the same time, each language model also has its own characteristics. The following is an introduction to each language model:

*3.2.1. N-Gram Model.* The N-gram model determines the occurrence of a word relative to the $n$ words that precede it. The biggest is when $n = 1$, that is, the occurrence of a word only interacts with the word itself [22]. This language model is called context-free model as shown in the following equation:

$$\begin{aligned} p(w_t|\text{Context}) &= p(w_t), \\ &= \frac{N_{wt}}{N}. \end{aligned} \tag{8}$$

When $n \geq 2$, it is called context-dependent model. In general application, it takes $n = 2$ or $n = 3$, that is, Bigram or Trigram. The advantage of N-gram model is that it takes into account the factor of the first $n - 1$ words, which have strong meaning in natural semantics. At the same time, using Bigram or Trigram can greatly simplify the calculation scale and improve the efficiency [23]. However, the n-gram language model itself has some limitations. For example, the N-gram model is based on the relationship between the corpus, the corpus is insufficient, and the training result is generally not ideal. Moreover, this model ignores the similarity relationship between words, only considers the relationship between words and context, but does not consider the relationship between words. Secondly, the N-gram model will have a statistical probability of 0 when some tuples have not appeared. This will lead to the probability of the whole language sequence to be 0. In this case, it often needs to be corrected to obtain accurate results.

*3.2.2. N-Pos Model.* The n-pos model is a language model derived from the N-gram model. The n-pos model is based on the following assumptions: considering the first $n$ words alone is not enough to represent the characteristics of the current word, and the word collocation in natural language is often determined according to the grammatical function of the word. Therefore, n-pos classifies the first $n$ words of

the current word according to the grammatical function, and these words determine the probability of the current word. This classification is called Part-Of-Speech, that is, the origin of the n-pos algorithm. The conditional probability formula of the n-pos model is shown in the following equation:

$$p(w) = p\left(w_1^T\right), \\ = \prod_{t=1}^{T} p\left(w_t | c\left(w_{t-n+1}\right), c\left(w_{t-n+2}\right), \ldots, c\left(w_{t-1}\right)\right). \tag{9}$$

$c$ is the part-of-speech mapping function, and $c(w_t)$ means to map the word $w_t$ to its part-of-speech category. If a language sequence has $T$ words and $K$ part-of-speech classifications, the conditional probability solution process of n-gram can be changed from $T^{n-1}$ to $K^{n-1}$, which greatly improves the calculation efficiency, and the improvement of this efficiency will not affect the decline of accuracy.

*3.2.3. Maximum Entropy Model.* The main idea of the maximum entropy model is as follows: when estimating the probability of random events, if the probability model satisfies certain constraints, then when the constraints are met, the unknowns are meaningless. In this case, the resulting distributions are usually similar, and the entropy of the received distribution is the largest [24]. The probability distribution formula of the maximum direct language model is shown in the following equation:

$$p\left(w_t | \text{Context}\right) = \frac{e^{\Sigma_i \lambda_i f_i(\text{context}, w)}}{Z(\text{context})}, \tag{10}$$

where $\lambda_i$ is the parameter and $Z(\text{context})$ is the normalization factor.

*3.3. Overall Framework Design of Text Tendency Analysis System.* This system is a text analysis system designed for the analysis of film review tendency. Its main process design consists of four parts: text preprocessing stage, text storage stage, text analysis stage, and result display stage. The processing flow chart is shown in Figure 2.

Firstly, the system obtains the comment information of the film as the initial corpus, then saves it as the training corpus after preprocessing the film information, and then enters the text analysis stage, including the training and classification of the model. The output is the trained text tendency classification model, shows the accuracy of the model through the display interface, and provides the interface to demonstrate the judgment of the tendency of the text [25]. According to the overall process of the system described above, the overall framework of the system is designed as shown in Figure 3.

The system is divided into three main modules: text preprocessing module, text storage module, and text analysis module. The following is the function introduction of each module:

(1) Text preprocessing module: the text preprocessing module is mainly responsible for text acquisition and processing. The text acquisition uses the customized spider crawler to obtain the text corpus. Compared with the general crawler, the indiscriminate crawling training corpus will produce a lot of noise for the extraction of text features. These noises will greatly reduce the accuracy of text training, thus affecting the final effect. Using customized crawlers, we can crawl the required corpus for analysis according to the characteristics of web pages. For example, for film comment information, the content to crawl includes comment subject, number of likes, comment scoring, comment label, and other information. The general crawler often simply removes the web page tag, leaving the text part as the training corpus. The granularity of such text corpus is very coarse, and the effect after word segmentation and filtering is often difficult to meet the requirements [26]. The differential classification of these information can not only remove the noise influence of the corpus but also facilitate the extraction of text features.

Another important function of the text preprocessing module is to segment and label the crawling corpus. For text tendency analysis, it is necessary to label the emotional words in the text. The tagging of emotional words is very helpful for the weight calculation of feature extraction. Emotional words are words stored in the emotional dictionary. The emotional dictionary often contains the part of speech, level, subjective, and objective attributes of these words. By judging the emotional words of the text, we can obtain the emotional level of the sentence in the text, and by judging the emotional level of the sentence, we can obtain the emotional level of the text segment, which is of great help to judge the emotional level of the whole text.

(2) Text enclosure: the text storage module saves the text into a special format and saves it into the distributed file system as the input of text analysis. In order to achieve fairness, the processed text needs to be divided into two categories: one is the training corpus with emotional level, and the other is the test corpus with implicit emotional level. Finally, the accuracy of the system is measured by comparing the scores obtained from the analysis of the test corpus with its actual scores.

(3) Text analysis module: the text analysis module uses two methods to represent text from the text. One is to vectorize the language according to standard neural network languages, and the other is to extract the content using grammar. Based on the text vector combined with word vector features and topics, the classification model is trained, and the classification model is finally obtained for analysis.

*3.3.1. Design of Text Preprocessing Module.* The text preprocessing module includes customized crawler module and word segmentation module, and its module composition diagram is shown in Figure 4.

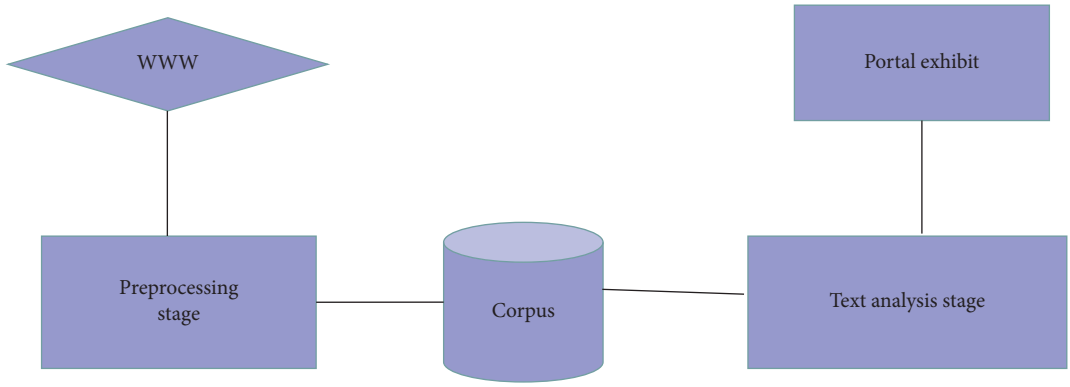


Figure 2: Schematic diagram of text analysis and processing flow.

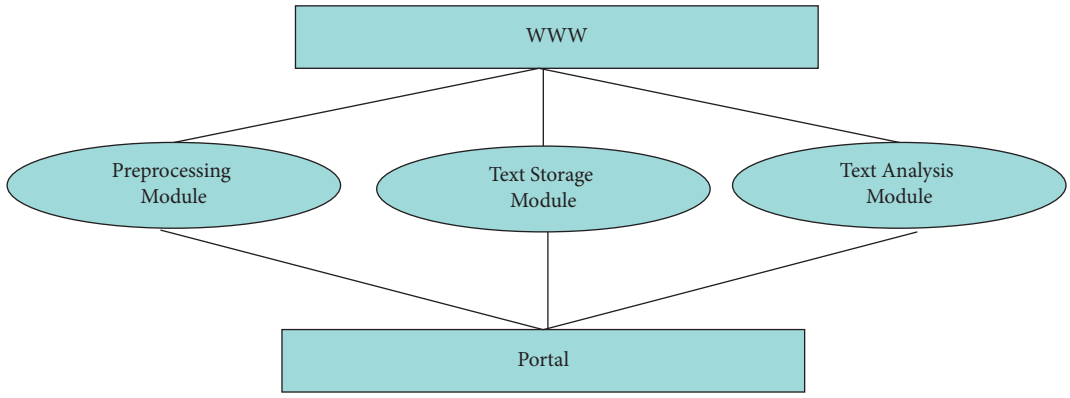


Figure 3: Overall framework of text tendency analysis system.

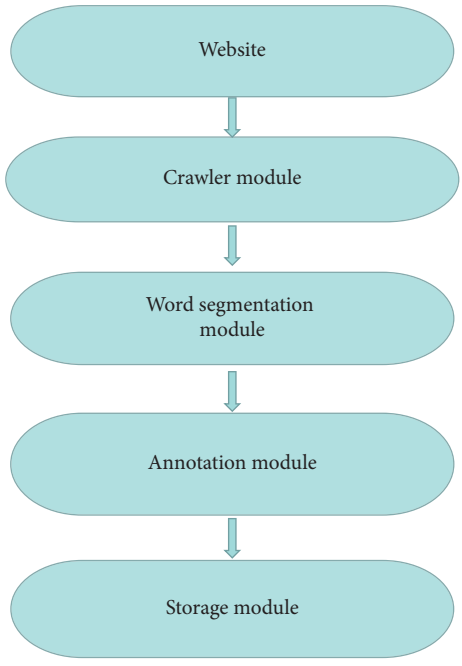


Figure 4: Text preprocessing flow chart.

*3.3.2. Design of Text Storage Module.* The main function of the text storage module is to store the crawled comment corpus and the preprocessed text. In order to prepare for text analysis, the system uses HDFS as the file system for text storage, saves the positive and negative text to HDFS, and saves the meta-information of the file in the database. When used to generate the model, the text analysis module reads the text meta-information from the database and then downloads the text to be trained from IDFS for local training. HDFS is generally used as the underlying storage system of the spark big data platform. Spark has a special interface to read text from HDFS and convert it into RDD. At the same time, RDD can also be persisted to HDFS as intermediate results. HDFS is a master-slave architecture. HDFS groups contain a personal name and some file nodes. As the owner of the node, the name node is mainly responsible for managing the names of system files and providing scheduling time for users to access data. Data nodes are storage nodes. Usually, each data node corresponds to a physical node in the cluster to manage the file data stored on it [27]. HDFS file system provides a name space and allows users to store data in files. Files are stored on a group of data nodes in the form of data blocks. Figure 5 is the architecture diagram of HDFS.

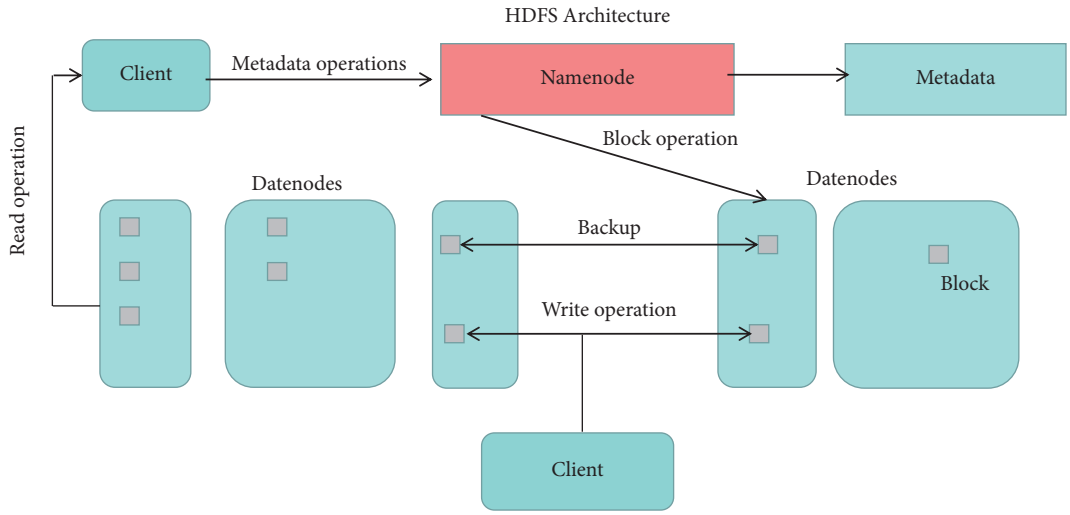


Figure 5: HDFS architecture.

The text storage module is a storage module designed based on HDFS. It adopts the way of small file storage. The crawler classifies the crawled comment text by movie category and saves it to the IDFS file system, then saves the path meta-information of the comment file on HDFS to the database, and saves other information of the file in the database, such as movie name, category, crawl time, and the number of comments. At the same time, the text preprocessing module will obtain the crawled comment information through the database and construct a corpus for text tendency analysis. The constructed corpus will be saved to the text storage system in a fixed directory structure. The text storage module is shown in Figure 6.

Take MySQL as the database information of comment text, and the database table storing comment information is shown in Table 1.

In the crawler crawling process, the text preprocessing module will process the saved comment text at the same time and save the processed comment text to the corpus. The corpus directory structure is organized according to certain rules. Its purpose is to provide the training module with available training corpus and use it to persist the trained model data, which is saved on HDFS.

*3.3.3. Design of Text Analysis Module.* The text analysis module is used to model the text. The text modeling method adopted by the system is to train the text model by integrating the text vector representation and text topic representation, which absorbs the advantages of both the language model based on statistics and the language model based on neural network. The module flow chart of the text analysis module is shown in Figure 7.

Among them, the Doc2vec algorithm is used for text vector representation, LDA Algorithm is used for text subject representation, and the Doc2vec algorithm is the vectorization of text fragments, which can be used to represent the characteristics of the text. The text topic representation adopts LDA Algorithm. LDA Algorithm is a text

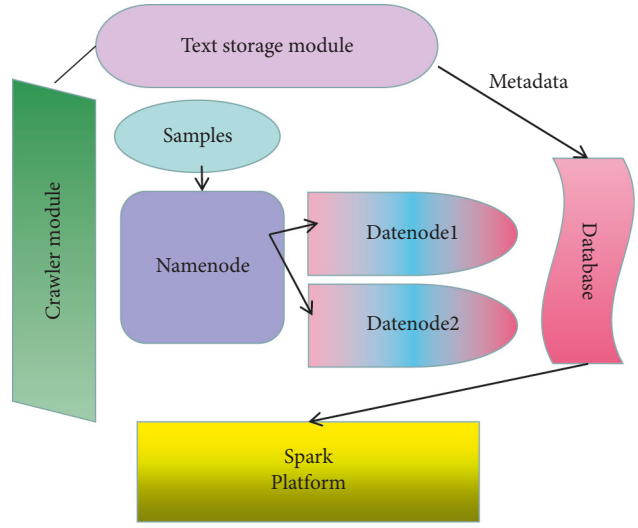


Figure 6: Structure diagram of text storage module.

topic model algorithm. It extracts the topic of the text, and sparse represents the text. It can also be used as the feature model of the text. The feature vector of the text can be obtained by fusing the two vectors, and then, the text classification model can be obtained by using the SGD classification model. This model is the language model finally used for propensity analysis.

## 4. Results and Analysis

### *4.1. Experimental Design and Analysis*

*4.1.1. Experimental Design and Process.* In the overall process design of the experiment, due to more preparations, the overall process is divided into two parts. The first part is the preparation process before text dimensionality reduction using the genetic algorithm, and the second part is the execution process of text feature dimensionality reduction using the genetic algorithm. The preparation process is shown in Figure 8.

Table 1: Comment information database.

| Column name | Type | Primary key | Is empty | Explanation |
|---|---|---|---|---|
| ID | Int | Yes | No | Primary key |
| MovieName | Varchar | No | No | Movie name |
| Type | Int | No | No | Movie types |
| CrawlDate | DateTime | No | No | Crawl time |
| AveScore | Int | No | No | Star |
| CommentsCount | Int | No | No | Number of comments |
| CommentsFilePath | Varchar | No | No | The comment text is in road on HDFS path |
| Tag | Int | No | No | Has been dealt with |

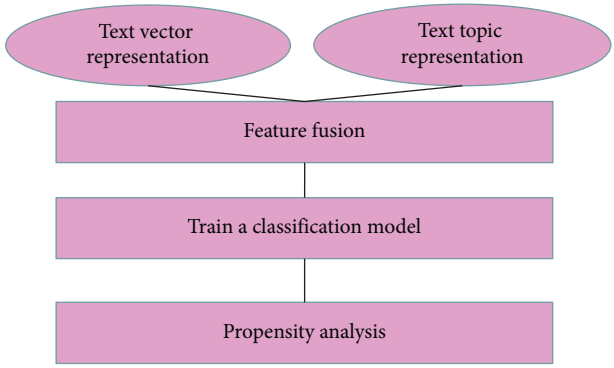


Figure 7: Structure diagram of text analysis module.

The specific operations in the preparation process are as follows:

The first thing to be decided is the classification algorithm used to generate the classification model. According to the characteristics of small data set, high text feature dimension, and secondary classification, the support vector machine with solid theoretical foundation is decided to be used as the classification algorithm of the training data generation model. We also include the preprocessing part of text data into the preparation process. In this step, we need to process punctuation, segment words, and remove stop words. The specific contents are as follows: firstly, the symbol table is made according to the common symbols and special symbols in the corpus, and the symbols in the text are accurately removed. Then, the precise mode in Jieba word segmentation introduced above is used to realize word segmentation and obtain the text feature set. Finally, the stop words in the text feature set are removed according to the stop word table of Harbin Institute of technology. Then, the text representation methods are selected. According to the comparative experimental analysis of three text representation methods in Chapter 3, the text data set used in the experiment has the best classification effect when using text Boolean representation method [28]. Therefore, the text Boolean representation method is selected to represent the text structurally. Finally, the text features need to be preliminarily filtered to obtain the individual gene group in the genetic algorithm because it is found that the classification effect is often poor when using full features to process text data. It shows that there are many redundant or irrelevant item features in the text feature set. Affected by these item features, the expected classification accuracy cannot be achieved. In order to improve the classification performance of the classification model and reduce the spatial complexity of the next genetic algorithm, a step is added before using the genetic algorithm. TF-IDF filtered feature selection dimensionality reduction algorithm is used to preliminarily filter the text features, and the first 10000 of the 19036 text features obtained after word segmentation are selected as the gene set of chromosomes.

The execution flow of dimensionality reduction of text features using the genetic algorithm is shown in Figure 9.

*4.1.2. Experimental Data and Parameters.* The data used in this experiment is the hotel psychiatric examination written by Professor Tan Songbo, with a total of 10,000 corpora, including 3,000 good corpus and 7,000 negative corpus. In order to reduce the influence of unreliable data distribution, 2000 positive and negative data are used in the training process, and then, 100 positive and negative corpora are selected from the corpus according to the experimental process. This experiment runs on Linux 413 system. The code language of the experiment is Python 3, and the classification algorithm is the support vector machine. Because in support vector machine, RBF kernel function is suitable for high-dimensional nonlinear classification, and the effect is the best in the comparison experiment with other kernel functions, RBF kernel function is adopted in the experiment, the penalty coefficient is set to 400, and gamma is $1/k$ ($k$ = the number of text features selected by individuals). In the parameter setting of the genetic algorithm, we set the number of individuals in the population to 20, the population termination algebra to 160 generations, the selection operator to RWS (roulette selection algorithm), the crossover probability to 0.7, and the mutation probability to 0.0001 as shown in Table 2.

*4.1.3. Experimental Results and Analysis.* As shown in Figure 10, the black line is the line graph of fitness function of the best individual in each generation of population, and the red line is the average line graph of fitness function of all individuals in each generation of population. From the broken line diagram, it can be found that both the optimal value and the mean value show a gradual upward trend in the early stage of population reproduction. We regard the difference between the two broken lines as the distribution of individuals in the solution space. It can be found that when using the traditional genetic algorithm to reduce the

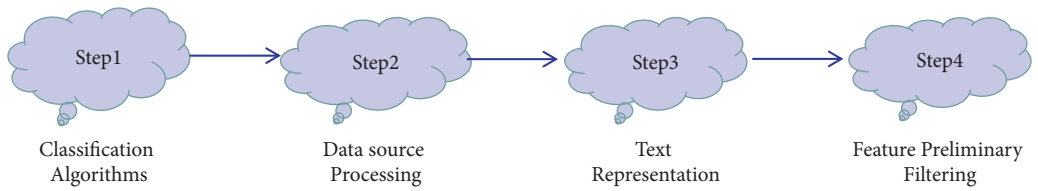


Figure 8: Preparation process.

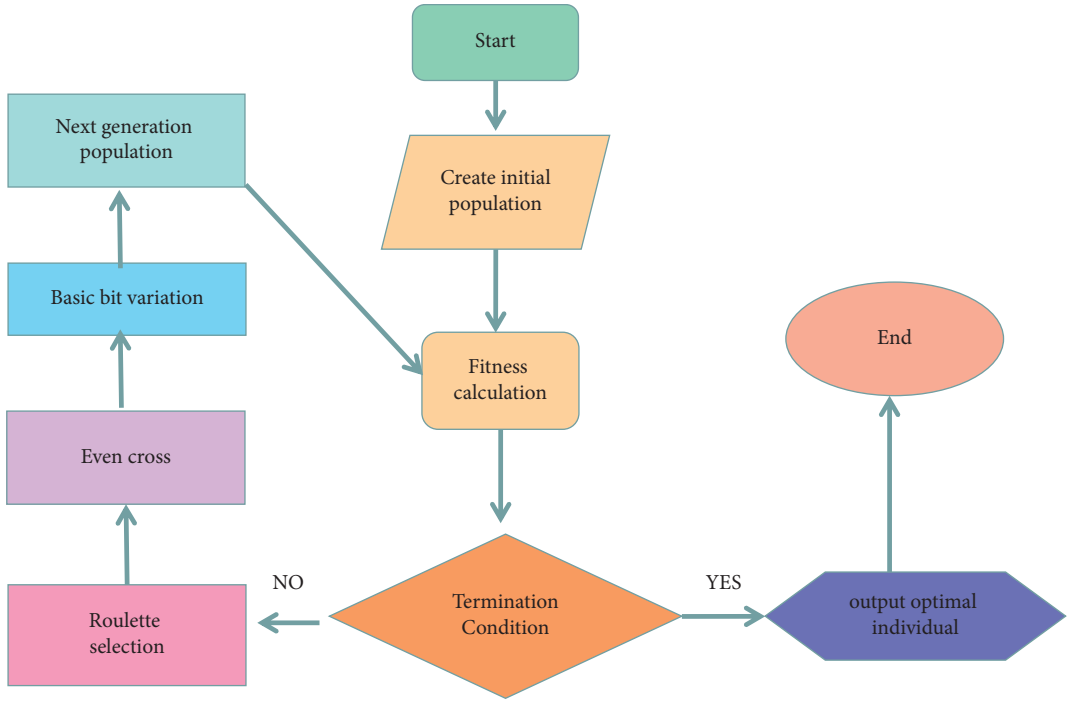


Figure 9: Execution algorithm flow.

Table 2: Experimental parameters.

| Parameter name | Parameter value |
|---|---|
| Kernel function | RBF |
| Penalty coefficient | 350 |
| Gamma | $1/d$ ($d$ = feature dimension) |
| Population size | 20 |
| Iteration number | 150 |
| Selection technique | RWS (roulette selection algorithm) |
| Crossover type | UC (even cross) |
| Crossover rate | 0.8 |
| Mutation type | SM (basic bit variation) |
| Mutation rate | 0.0002 |

dimension of text features, the distribution of individual solutions formed by different text feature subsets in the solution space changes constantly between dispersion and concentration, and the distribution of individual solutions in the population is relatively scattered in the periods of 1 to 35 generations and 93 to 125 generations. During the 36–92 generation and 126–160 generation, the distribution of individual solutions is more concentrated. The dispersion after each concentration is a process of jumping out of the local optimum. Under the condition that the maximum reproduction algebra is 160, we can regard the last 126 to 160

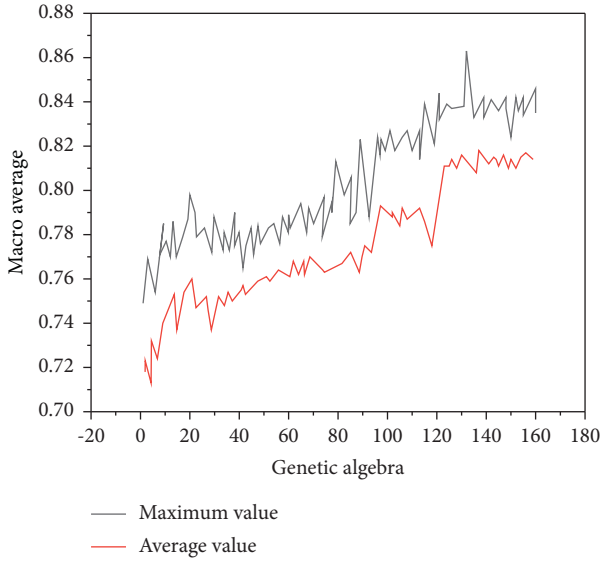


Figure 10: Population performance of different generations.

generations as the last population convergence part. In the last population convergence part, with the help of the green broken line, we can easily find that the individual with the maximum fitness function value in the reproduction process

appears in the 132 generation population, and this individual is the optimal solution we are looking for. Comparing it with the text feature subset obtained by the filter algorithm TF-IDF and CHI at the optimal macro average value, it can be found that the optimal individual obtained by the genetic algorithm has better recall rate, accuracy, and macro average value than a single filter algorithm, in which the recall rate is significantly improved.

Table 3 shows the comparison of recall rate, accuracy rate, macro average value, and characteristic dimension when the macro average value is obtained based on CHI, TF-IDF, and GA. It can be found that GA obtains the highest macro average value when the dimension is 5648, CHI obtains the highest macro average value when the dimension is 8000, and TF-IDF obtains the highest macro average value when the dimension is 9000 (CHI and TF-IDF obtain the highest macro average value through multiple comparisons by dichotomy). The experimental results show that the highest macro average value of CHI is similar to that of GA, the macro average values of both are higher than those obtained by TF-IDF method, and the recall rate of GA is significantly higher than that of CHI and TF-IDF of filter algorithm. Compared with the dimensionality reduction effect at this time, it is obvious that the dimensionality reduction effect of GA is much better than that of CHI and TF-IDF. As shown in Figure 11, the optimal dimension numbers of GA, TF-IDF, and CHI are 5648, 9000, and 8000, respectively. Compared with the dimension number 19033 after word segmentation, the dimension reduction rates of GA are 70.3%, 52.7%, and 58%, respectively. The dimension reduction rates of GA are 17.6% and 11.7% higher than those of TF-IDF and CHI, respectively. The dimension reduction effect is remarkable. By dividing the experiment from the hotel review data, the experiment shows that this paper is based on the CHI filtering algorithm, TF-IDF filtering algorithm, and feature selection algorithm based on genetics and the text feature selection algorithm based on genetic algorithm. In the case of macroscopic media (such as distribution effects), the dimensionality reduction ability is the best, and the distribution efficiency is good.

*4.2. Research and Analysis of Text Feature Dimensionality Reduction Based on Improved Genetic Algorithm.* In order to solve the problems of the TF-IDF algorithm and enhance the class discrimination ability of the algorithm, the mutual information filtering feature selection algorithm is introduced to calculate the text feature weight, and the traditional TF-IDF algorithm is improved. Through mutual information algorithm, we can get the information weight value between feature words and single category documents. It is planned to replace the calculation significance of the number of documents with feature words in TF-IDF with the dispersion of feature words and the weight value of each category information. Formula (11) calculates the average value of mutual information of feature word $a$ among various texts.

Table 3: Comparison of GA, CHI, and TF-IDF when macro average is the best.

| | Precision (%) | Recall (%) | Macro-F (%) | Dimension |
|---|---|---|---|---|
| GA | 85.50 | 93.29 | 87.30 | 5545 |
| CHI | 85.50 | 87.07 | 87.3 | 8220 |
| TF-IDF | 85 | 87.87 | 88.9 | 9560 |

$$\overline{MI} = \frac{1}{j}\sum_{1}^{j} MI\left(a, c_j\right). \tag{11}$$

In formula (12), by calculating the standard deviation of the mutual information weight between the feature word and various texts, the dispersion degree of the weight value of the feature word and each category of information is expressed.

$$s = \sqrt{\frac{1}{j}\sum_{1}^{j}\left(MI\left(a, c_j\right) - \overline{MI}\right)^2}. \tag{12}$$

Finally, based on TF-IDF and MI, a new text feature weight calculation method with the overall contribution ability and category contribution ability of text features is obtained as shown in the following formula:

$$\mathrm{FUSION}_a = \frac{N_{a,A}}{N_A}\log\left(1 + e^s\right). \tag{13}$$

In order to verify that the text features selected by the filtering selection algorithm of multirule fusion have better classification effect, the following experiments are carried out. TF-IDF, MI, and fusion algorithms are used to calculate the weight value of the experimental data, respectively, and the top 10000 text features with high weight value are filtered. The text Boolean representation method is used for text representation, and support vector machine is used as the classification algorithm to generate the model for experimental effect comparison. The data used in the experiment is the same as the experimental data used in Chapter 4. It is the hotel comment emotion analysis corpus collected by Professor Tan Songbo. There are 4000 training sets and 200 test sets. The kernel function of support vector machine uses RBF, the penalty coefficient is set to 400, and the gamma is 1/10000.

*4.2.1. Experimental Results.* The experimental results are shown in Table 4. The first 10000 features are obtained by using the new filtering algorithm of multirule fusion. The effect of the model after training on the test set is better than that of MI and TF-DF. The accuracy, recall, and macro average are more than 80%. Compared with MI and TF-IDF, the accuracy, recall, and macro average are 8% and 3% higher, respectively, 5% and 4% higher, respectively; and 5.9% and 2.9% higher, respectively. The feasibility of the algorithm in the experimental data is proved. Therefore, in the preparation process, the multirule fusion filtered feature

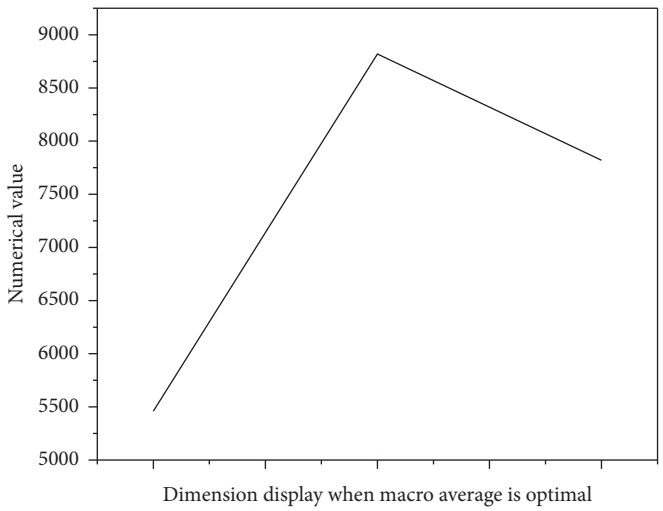


Figure 11: Best dimension display.

Table 4: Effect comparison of gene group selection algorithm.

| | Precision (%) | Recall (%) | Macro-F (%) |
|---|---|---|---|
| MI | 70 | 73 | 76 |
| TF-IDF | 75 | 77 | 75.4 |
| FUSION | 80 | 81 | 81.9 |

selection algorithm is used to filter the text features to form the individual gene group to be selected.

*4.2.2. Design and Process.* In order to improve the text feature selection ability of the genetic algorithm, improve the convergence speed, and achieve the effect of dimensionality reduction, the preparation process and execution process of text feature dimensionality reduction using the traditional genetic algorithm in the previous chapter are improved step by step. The overall flow chart of the improved algorithm is shown in Figure 12.

In the preparation process stage, in order to make the text features obtained after preliminary filtering have better performance ability and reduce the omission of excellent text features, TF-IDF algorithm is improved. According to the problems of the TF-IDF algorithm and the advantages of the mutual information algorithm, the two are fused to form a new multirule fusion filtering feature selection algorithm, which makes the text feature gene group obtained after preliminary filtering have better expressiveness and persuasion.

*4.2.3. Experimental Analysis.* Figure 13 is a broken line diagram of individual performance of different generations in the improved genetic algorithm, in which the black broken line represents the maximum value of individual fitness function in each generation, and the red broken line represents the average value of all individual fitness functions in each generation. It can be found from the red broken line that when the population continues to multiply, the

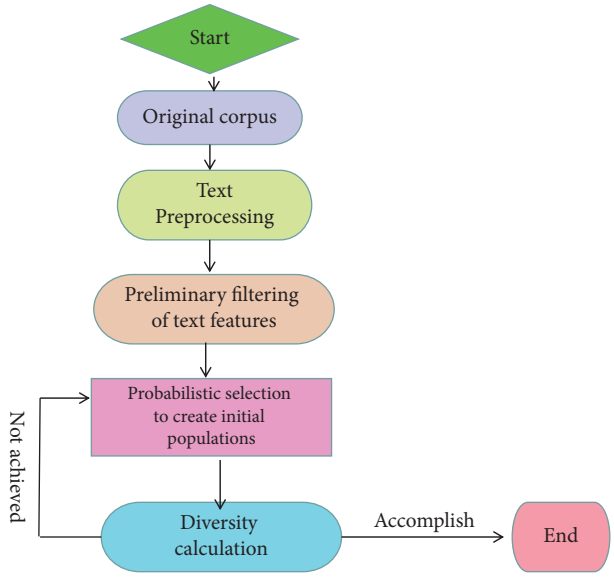


Figure 12: Improved algorithm flow.

individual's external performance is constantly improving, and the improvement is relatively gentle most of the time. Measuring the difference between the black broken line and the red broken line represents the dispersion degree of the individual population. It can be found that the general difference within the stage range is gradually shrinking, which is in line with our expectations for population reproduction, that is, the population gradually converges in the reproduction process. When the population reproduces to 86 generations, the optimal individual appears, and the fitness function value is 0.8925.

Figure 14 is a comparison diagram of the energy function of the genetic algorithm and the genetic improvement algorithm. By comparing the average physical activity cost of each population, it can be seen that the genetic algorithm and the improved algorithm in this form of physical activity

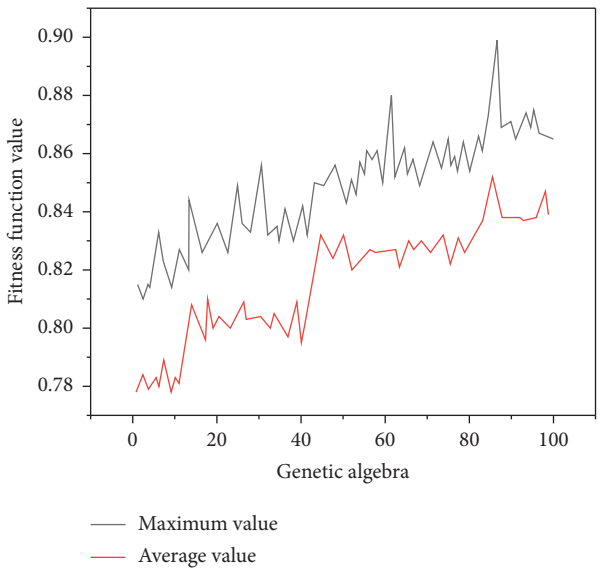


Figure 13: Individual performance of improved genetic algorithm.

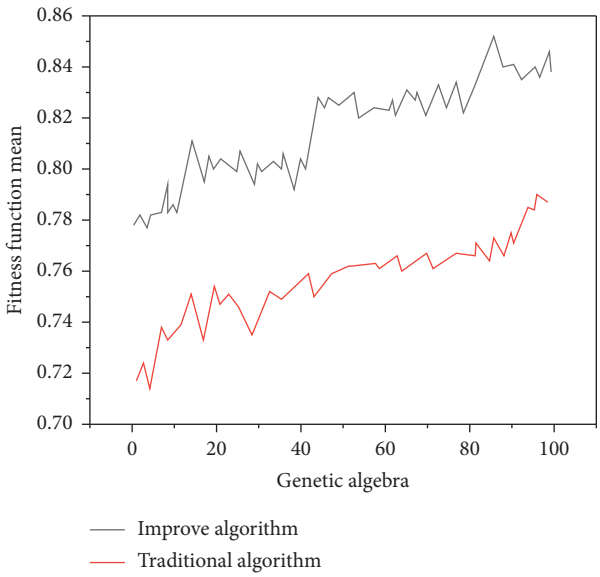


Figure 14: Comparison of fitness function between simple genetic algorithm and improved genetic algorithm.

usually show a significant increase in the growth rate of 100 generations. In the first stage of population breeding, all the benefits of genetic improvement algorithms over traditional genetic algorithms are the result of many different fusion algorithms and methods that select workers in repair, and therefore improve the performance of individuals in the first population. In the genetic algorithm, the energy function increases gradually, and the increase is not obvious, and the frequency of the population mean decreases, while in the improved genetic algorithm, the population mean increases by 3 times. It can be seen that the improved genetic algorithm is better than the traditional genetic algorithm.

The experimental results show that the features screened by the improved genetic algorithm have better feature screening ability, larger dimension reduction range and higher accuracy, recall, and macro average than the single TF-IDF algorithm and CH algorithm. It is faster than the traditional genetic algorithm and can find the optimal individual with better dimension reduction effect and classification performance in a shorter time.

## 5. Conclusion

The development of the Internet is more and more rapid. With the continuous progress of science and technology, mankind will slowly enter the intelligent era. A series of applications based on text analysis will gradually affect people's way of life. However, the field of text analysis is still under exploration, and the old research methods are no longer suitable for the current needs. Based on the latest research results and combined with the genetic algorithm, this paper makes a systematic research on the field of text analysis. In this paper, the experimental process is clarified in a phased way. After the relevant operations of the preparation process, the dimension of text features is reduced by the genetic algorithm, so that the dimension-reduced text features can achieve better results in the application of text classification, and is based on the traditional genetic algorithm. According to the problems of convergence speed and local optimization, the relevant steps are improved, and the effectiveness of the improvement is proved by experiments. The main improvements include the following two aspects:

(1) Improvement in the preparation process stage: the filter feature selection algorithm based on the multirule fusion is used to select gene groups, which avoids the problems of the TF-IDF filter feature selection algorithm in gene group selection, and then, the rationality of the improvement is verified by classification performance evaluation.

(2) Improvement of execution process stage: the generation mode of individuals in the initial population is changed, the operation of population diversity calculation is increased, and the fitness function is modified to increase the influence ability of dimension. Finally, in order to speed up the convergence speed, the probability of the constant crossover operator and variation operator is changed into an adaptive mode. The experimental results show that the improved genetic algorithm improves the convergence speed and local optimization.

## Data Availability

The labeled data set used to support the findings of this study is available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

This work was supported by the School of Foreign Languages and Literature, Tianjin University.

## References

[1] A. A. Liednikova, D. V. Shypik, and P. I. Bidyuk, "Project risk analysis using text data mining of comments in project management system jira," *System Research and Information Technologies*, vol. 2020, no. 2, pp. 121–136, 2020.

[2] Z. Li, K. Guo, M. Liao, A. Zhao, M. Tian, and Y. Wang, "Micro-hybrid energy storage system capacity based on genetic algorithm optimization configuration research," *International Core Journal of Engineering*, vol. 6, no. 2, pp. 78–83, 2020.

[3] C. C. Hsieh and J. Z. Chiu, "System review: a text analysis on supply chain finance," *Universal Journal of Management*, vol. 8, no. 2, pp. 29–32, 2020.

[4] T. Rizvi, "Identification of mind-set of students through web based basic psychological text and graphical analysis system," *International Journal of Management and Humanities*, vol. 4, no. 7, pp. 33–36, 2020.

[5] P. Wang, Y. Zhang, and H. Yang, "Research on economic optimization of microgrid cluster based on chaos sparrow search algorithm," *Computational Intelligence and Neuroscience*, vol. 2021, no. 3, 18 pages, Article ID 5556780, 2021.

[6] H. Liu, H. Liu, X. Wang, W. Shao, X. Wang, and J. Du, "Smartmeeting: an novel mobile voice meeting minutes generation and analysis system," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 521–536, 2020.

[7] H. Bai, H. Yu, G. Yu, L. Rocha, and X. Huang, "Analysis on an auto increment detection system of Chinese disaster weibo text," *Journal of Universal Computer Science*, vol. 27, no. 2, pp. 230–252, 2021.

[8] O. I. Sheluhin, D. V. Kostin, and M. G. Gorodnichev, "Multiclass classification of anomalous states of computer systems by means of intellectual analysis of system journals," *Automatic Control and Computer Sciences*, vol. 54, no. 6, pp. 549–559, 2021.

[9] E. S. Izrailova and A. S. Badaeva, "Analysis of the speech signal quality of the Chechen speech synthesis system," *Automatic Documentation and Mathematical Linguistics*, vol. 55, no. 2, pp. 74–78, 2021.

[10] M. D. Mufti, M. Y. Zargar, and A. W. Kumar, "Adaptive predictive control of flywheel storage for transient stability enhancement of a wind penetrated power system," *International Journal of Energy Research*, vol. 46, no. 5, pp. 6654–6671, 2022.

[11] Y. Meng, Y. Liang, Q. Zhao, and J. Qin, "Research on torsional property of body-in-white based on square box model and multiobjective genetic algorithm," *Mathematical Problems in Engineering*, vol. 2021, no. 41, pp. 1–13, Article ID 7826496, 2021.

[12] M. Tominaga, M. Okajima, M. Yamagishi, M. Shinagawa, J. Katsuyama, Y. Matsumoto et al., "Noise analysis of an electro-optic sensor system," *Optical Review*, vol. 28, no. 6, pp. 704–715, 2021.

[13] M. Yue and X. Wang, "Research on control strategy of ship energy management system based on hybrid ga and pso," *International Core Journal of Engineering*, vol. 6, no. 5, pp. 185–193, 2020.

[14] R. Shi, "Research on data mining system based on artificial intelligence and improved genetic algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 8, pp. 1–12, 2020.

[15] Y. M. Yeshilbashian, A. A. Asatryan, and T. G. Ghukasyan, "Plagiarism detection in Armenian texts using intrinsic stylometric analysis," *Proceedings of the Institute for System Programming of RAS*, vol. 33, no. 1, pp. 209–224, 2021.

[16] D. M. Karshiyevich and H. H. Syed, "Basic principles of creating software system to control and correct errors in text," *SSRN Electronic Journal*, vol. 7, no. 11, pp. 8–14, 2020.

[17] Y. Chen, H. Jiao, H. Zhou, J. Zheng, and T. Pu, "Security analysis of qam quantum-noise randomized cipher system," *IEEE Photonics Journal*, vol. 12, no. 99, p. 1, 2020.

[18] Z. Kafadar, "A geophone-based and low-cost data acquisition and analysis system designed for microtremor measurements," *Geoscientific Instrumentation Methods and Data Systems*, vol. 9, no. 1, pp. 365–373, 2020.

[19] B. J. Galli and L. C. Wood, "How to apply system analysis and system thinking to lean six sigma initiatives," *International Journal of Service Science, Management, Engineering, and Technology*, vol. 12, no. 4, pp. 1–25, 2021.

[20] A. Chaudhuri, N. Sinhababu, M. Sarma, and D. Samanta, "Hidden features identification for designing an efficient research article recommendation system," *International Journal on Digital Libraries*, vol. 22, no. 6, pp. 1–17, 2021.

[21] P. Chang and H. J. Tsai, "Text-image complementarity and genre in English as foreign language textbooks," *Semiotica*, vol. 2022, no. 244, pp. 53–80, 2022.

[22] O. Agarwal, Y. Yang, B. C. Wallace, and A. Nenkova, "Interpretability analysis for named entity recognition to understand system predictions and how they can improve," *Computational Linguistics*, vol. 47, no. 1, pp. 1–24, 2021.

[23] P. J. Younse, J. E. Cameron, and T. H. Bradley, "Comparative analysis of model-based and traditional systems engineering approaches for arCHItecting a robotic space system through automatic information transfer," *IEEE Access*, vol. 9, no. 99, p. 1, 2021.

[24] Z. N. Al-Kateeb and M. Jader, "Encryption and hiding text using dna coding and hyperchaotic system," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 19, no. 2, pp. 766–774, 2020.

[25] A. L. Mahmood, A. M. Shakir, and B. A. Numan, "Design and performance analysis of stand-alone pv system at al-nahrain university, baghdad, Iraq," *International Journal of Power Electronics and Drive Systems*, vol. 11, no. 2, pp. 921–930, 2020.

[26] F. N. Al-Wesabi, S. Alzahrani, F. Alyarimi, M. Abdul, and M. M. Almazah, "A reliable nlp scheme for English text watermarking based on contents interrelationship," *Computer Systems Science and Engineering*, vol. 37, no. 3, pp. 297–311, 2021.

[27] C. Payant and P. Bell, "Very easy, it's an English class, therefore they should not rely on a French text: English language teachers' beliefs regarding l1 use for literacy instruction," *Language TeaCHIng for Young Learners*, vol. 4, no. 1, pp. 143–170, 2022.

[28] A. Kusuma, M. H. Santosa, and I. Myartawan, "Exploring the influence of blended learning method in English recount text writing for senior high school students," *Jet*, vol. 6, no. 3, pp. 193–203, 2020.