Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8048335, 14 pages
https://doi.org/10.1155/2022/8048335

*Research Article*

# Naive Bayesian Prediction of Japanese Annotated Corpus for Textual Semantic Word Formation Classification

**Zhoushao Hao**

*Luoyang Normal University, Luoyang Henan 471934, China*

Correspondence should be addressed to Zhoushao Hao; haozs@lynu.edu.cn

Received 13 January 2022; Revised 11 February 2022; Accepted 19 February 2022; Published 16 March 2022

Academic Editor: Gengxin Sun



With the rapid development of Japanese information processing technology, problems such as polysemy and ambiguity at the text and dialogue level, as well as unregistered words, have become increasingly prominent because computers cannot fully "understand" the semantics of words. How to make the computer "understand" the semantics of words accurately requires the computer to "understand" the rules of converting and integrating words into words from the perspective of semantics. Traditional Japanese text classification mostly adopts the text representation method of vector space model, which has the problem of confusing classification effect. Therefore, this paper proposes the topic of constructing a semantic word formation pattern prediction model based on a large-scale annotated corpus. This paper proposes a solution that combines Japanese semantic word formation rules with pattern recognition algorithms. Aiming at this scheme, a variety of pattern recognition algorithms were compared and analyzed, and the naive Bayesian model was decided to predict semantic word formation patterns. This paper further improves the accuracy of computer prediction of Japanese semantic word formation patterns by adding part of speech. Before modeling, the parts of speech of words are automatically tagged and manually checked based on the original annotated corpus. In the research on predicting Japanese semantic word formation patterns, this paper builds a semantic word formation pattern prediction model based on Naive Bayes and conducts simulation experiments. We divide the eight types of semantic word formation patterns in the annotated corpus into two groups, and divide the obtained sample sets into training sets and test sets, so that the Naive Bayes model first learns semantic word formation rules based on the training sets of each group. Semantic word formation patterns are predicted on the test set for each group. The simulation results show that the prediction model of semantic word formation mode has a generally high degree of fit and prediction accuracy. The prediction model of semantic word formation pattern based on this theory can ensure that the computer can judge the semantic word formation pattern more accurately.

## 1. Introduction

With the rapid development of information technology, a large amount of Japanese text data is generated every moment on the Internet. Traditional manual classification methods can no longer meet the needs of society, so fast and efficient automatic Japanese text classification technology has become a hot research topic [1]. Although Japanese text classification technology is widely used in spam filtering, search engines, and information management, and has achieved rapid development, the actual classification performance is still relatively low, and there is still a lot of room for improvement in classification accuracy and efficiency. In these massive Japanese text data, a lot of valuable information is contained in them, and people need to actively explore and explore. In order to deal with this situation and obtain valuable information from massive Japanese text data in time, Japanese text classification technology came into being. Japanese text classification technology plays an important role in research fields such as information organization and Japanese text mining. It can effectively help people extract needed information from disordered Japanese text data. It is a powerful method of Japanese text processing technology [2–5]. Because of the fast and efficient processing efficiency of Japanese text classification technology, it has been widely used in information retrieval, search engines, and spam filtering.

At present, in text semantic research at the text level, most scholars focus on introducing Japanese text classification methods into sentiment classification. However, due to the various ways of expressing emotions in Japanese texts, the semantic information in Japanese texts is very important for understanding emotional expressions. Therefore, obtaining this semantic information is very necessary for the recognition of emotional tendencies [6–8]. With the continuous popularization of Web 2.0, more and more posts are actively published by ordinary texts, such as blogs, various comments, forum posts, and so on. Although people can easily obtain this information through the Internet, people cannot obtain any valuable information from the massive amounts of data in the Internet if they are not summarized and sorted out. Unprocessed data is just a bunch of meaningless symbols. Only by analyzing and extracting valuable information can it be used by people. How to extract this information by effective means is a current research hotspot in the computer field. The development of industry and the Internet has spawned many application scenarios that require machines to perform emotion classification, such as scoring prediction. This research first reviews the current status of opinion mining research in natural language processing, and then discusses the advantages and disadvantages of traditional content recommendation algorithms in content recommendation and rating prediction, and makes a feasible algorithm for recommendation and rating algorithms with Japanese text comments as the data source. After the analysis, it was found that the naive Bayes Japanese text scoring prediction algorithm based on the topic model is very suitable for the Japanese text scoring prediction problem [9–11].

This paper proposes a convolutional naive Bayes parallel classification model based on semantic expansion. Since the web short Japanese text data set has the characteristics of fuzzy semantics and sparse features, the method of constructing topic-feature two-tuples is used to achieve the purpose of semantic expansion of Japanese text features, and the two-tuples are used as the Bayesian classification model. We use the convolutional Naive Bayes classification model to further optimize the data features, and use the Softmax function to classify; then combine the MapReduce framework in the process of constructing feature two-tuples and parameter training, respectively, in the data preprocessing and the parameters of the classification model tuning two parts to complete the parallel design. It is verified by design experiments that the convolutional naive Bayes classification model based on semantic expansion improves the accuracy and classification efficiency of the classification model when processing web short Japanese text data. In order to solve these problems, this paper attempts to introduce ontology, using ontology hierarchical structure and attribute constraints to match keywords with domain ontology concepts, and establish a concept vector space model for Japanese text classification. It aims to solve the multisense and conceptual hierarchical problems in Japanese text classification, overcome the shortcomings of keyword-based classification methods, and improve the accuracy of classification [12–14]. At the same time, this paper also studies the relationship between Japanese text classification and personalized information retrieval, analyzes the text interest model, and proposes a text interest model establishment and adjustment algorithm to make the classification result more in line with the text intent.

## 2. Related Work

In recent years, due to the development of computer technology and the improvement of computing and storage capabilities, Japanese text classification has gradually begun to use convolution kernel operations, a naive Bayes algorithm that takes up a lot of machine resources. When using a machine learning model to process Japanese text classification problems, the Japanese text data is divided into a test set and a training set according to a certain ratio. The classifier learns through the training data, and gradually optimizes the model parameters to make the classification effect better and better, and finally get a classification model. The application of machine learning in Japanese text classification has greatly improved the efficiency and accuracy of classification. Based on the above improvements, the Japanese text classification system proposed in this article was formed. Experiments were conducted on a large number of labeled news data sets. The improved hybrid naive Bayes model proposed in this article was combined with the traditional machine learning model SVM and Naive Bayes. The comparison of the classification performances of the others verifies that the improved hybrid naive Bayes model has a better classification accuracy [15–17].

Considering the particularity of Japanese text, Ma et al. [18] proposed a word vector and character vector training model based on Japanese kanji information and radical information, using a radical conversion mechanism to allow words with similar semantics to be mutually in the vector space. At the same time, the word vector is discarded, and the word vector is used as the Japanese text input, which can better control new words and rare words. Aiming at the problem of one-time ambiguity in Japanese texts, Zhang et al. [19] proposed an improved model called "Topic-SG" to realize the calculation of topic-word vectors, and merge the word vectors and topic vectors to a certain extent. The polysemous words frequently appearing in Japanese have a special influence on short Japanese texts. In the field of academic research, Bayesian algorithm is constantly improving and innovating. Horvat [20] proposed Bayesian association rules, which combine Bayesian network with association rules, which can accurately assess the conditional relevance and independence between item sets. We combine the naive Bayes classifier with Bayesian network, and then apply it to network intrusion detection, and have achieved certain improvements. In order to solve the problem of accurate probabilistic inference of any Bayesian network, because the time consumed in a single-machine serial environment is relatively large, so MapReduce is used to convert the Bayesian network model into tasks that are executed in parallel. The $K$ neighbor algorithm is also very simple and efficient, but when the training data set is very large, the amount of calculation will increase very much.

Naive Bayes is based on the following assumption: each feature word in the document is independent of each other. Although this violates the rules of natural language, after IR conversion, the experimental results show that the Bayesian algorithm is quite accurate and simple and easy to implement.

Zhuo et al. [21] proposed a smooth and naive Bayes model that combines the characteristics of word independence and word independence. Previously, a more widely used confidence estimation model used word graphs to calculate the posterior probability of words to achieve detection and recognition. Compared with the self-confidence estimation model, the effect of this model is very obvious for the wrong words in the sentence. Cheng et al. [22] proposed a new Bayesian network learning algorithm in the context of big data, which integrates MMHC, TPDA, and REC. It consists of three stages: data preprocessing, individual overall learning, and concentrated overall learning. The three-stage algorithm can efficiently learn Bayesian algorithm from big data, and has higher accuracy than a single MMHC, TPDA, and REC. Aiming at a series of shortcomings such as the long time-consuming training and testing process of the existing large-scale Japanese text document classification on a single machine, a parallel Bayesian Japanese text classification algorithm based on the MapReduce architecture was designed, it is close to linear acceleration ratio [23–25]. In response to the threat of botnets, a Map Reduce Bayesian algorithm based on the hadoop platform is proposed. This method takes the host as the analysis object, extracts the characteristics of the communication traffic between the two hosts and uses it as the input of the Bayesian classification algorithm, and calculates the prior and conditional probabilities in the Bayesian algorithm training phase in parallel to form Bayesian classification. It can learn to recognize the traffic of botnets, and use the Bayesian classifier formed in the training stage and the posterior probability of parallel calculation in the detection stage to realize the detection of botnets. Through experimental comparison, it can be found that for large-scale Japanese text data, the accuracy of the deep learning model is better than that of the traditional machine learning model. This may be because the deep learning model is better at extracting complex and multidimensional features that are difficult to explain. At the same time, while extracting local features, the convolutional naive Bayes model ignores contextual semantic contact information. The addition of the two-way long and short-term memory model LSTM can effectively improve this problem and improve the accuracy of classification. Finally, although the accuracy of the model proposed in this paper is higher than that of a single Bayesian model, due to the addition of the TF-IDF value as the weight calculation and the LSTM model, the amount of calculation is greatly increased and the running time is increased.

## 3. The Semantic Word Formation Mode of Japanese Text with the Aid of Corpus

*3.1. Corpus Level Classification.* Corpus information gain is a relatively common algorithm in the process of text feature selection. Because of its simple probability calculation, it is widely used in actual feature selection tasks. It measures the amount of information carried by the feature word by calculating the difference in entropy before and after the presence or absence of a certain feature word. The larger the information gain value, the more information the feature carries. Whether the selection of the training document set is appropriate has a greater impact on the performance of the document classifier. The training document set should be able to broadly represent the documents in each document category that the classification system needs to process. Generally speaking, the training document set should be a recognized artificially classified corpus.

$$\text{if}\{(i,j) \longrightarrow (i',j'), k \in [i,j]\}, \text{ s.t.}\,[L_k,k] \subseteq [1,2,\ldots,n]. \tag{1}$$

Japanese text classification is a guided learning process. It finds the relationship model between document features and document category based on a set of marked training documents, and then uses the learned relationship model to make category judgments on new documents. The document classification process can be described more formally. Suppose there is a set of document concept class C and a set of training documents D. The document concept class and the documents in the document library may satisfy a certain conceptual hierarchical relationship $h$. That is, what kind of language elements (document features) are used and what mathematical forms are used to organize these features to represent the document. This is an important technical issue in document classification. The current Japanese text classification methods and systems mostly use words or phrases as language elements that represent the semantics of documents, and the representation models are mainly vector space models.

$$\sum f(k-1) \times f(k) - \sum f(n-1) \times \frac{k-1}{n-1} = 0, \quad k \subseteq [1,n]. \tag{2}$$

Language is an open system, as a kind of written materialized or electronic document of language is also open. Its size, structure, language elements and information contained are all open, so its characteristics are also unlimited. The Japanese text classification system should select as few and accurate document features as possible and closely related to the concept of the document subject for Japanese text classification. Among the various feature selection methods, the calculation of the DF method is the simplest, and it is also one of the most effective Japanese text feature selection methods. However, due to the lack of the necessary theoretical foundation, document frequency has always been regarded as a stopgap measure to improve the efficiency of Japanese text classifiers, and cannot be regarded as a method for selecting Japanese text features with strict entropy.

*3.2. Naive Bayes Algorithm.* Naive Bayesian (NB) is a classification algorithm based on Bayes' theorem and the assumption of independence of feature conditions. Compared with other classification algorithms, the Naive Bayesian

algorithm is simpler, has a solid mathematical foundation, and has more small error rate. The basic idea is to calculate which category the data belongs to according to the prior probability and conditional probability of each feature for a data that needs to be classified.

$$\sum \frac{f\left(\left|L_i \times L_{k-1}\right| - k\right)}{\left|L_i \times L_{k-1}\right|} - f(k) - \sum_{i=1}^{n} \frac{\left|L_i \times L_{k-1}\right|}{\left|L_i \times L_{k+1}\right|^{W}} = 0. \tag{3}$$

According to the conclusion of the comparative experiment in the literature, IG and Z2 have the best statistical effects among these methods, and can achieve a high compression rate (only 2% features are retained) without loss of classification accuracy. If about 10% of the features are retained, the effect of the DF method can be compared with that of the lG CHI. In the case of too much calculation, the DF method can be used instead of lG Z2 statistics to achieve a good balance between accuracy and efficiency. The basic idea of the algorithm is that for a document $d$ to be classified, the system finds the $k$ most similar training documents in the training set through similarity.

$$\begin{cases} f(0) = 0, \\ f(i) - f\left(\left|L_i L_{k-1}\right|\right) = \lim\limits_{i \longrightarrow \infty} f(i,k), \\ f(n) = 1. \end{cases} \tag{4}$$

On this basis, each document category is scored, and the score is the sum of the similarity between the documents belonging to this category in the $k$ training documents and the test documents. That is to say, if there are multiple documents belonging to one category among the $k$ documents, the score of this category is the sum of the similarity between these documents and the test document. After the scores of the categories of these $k$ documents are counted, they are sorted according to the scores. Fisher discriminant analysis is used to reduce the dimensionality of the original data and extract discriminative features. A threshold should also be selected, and only categories whose score exceeds the threshold will be considered. The test documents in Table 1 belong to all categories that exceed the threshold.

The solid dots and the hollow dots represent two types of samples. H is the classification line. H1 and H2 are the lines that pass the closest samples to the classification line and are parallel to the classification line. The distance between them is called the classification interval. The so-called optimal classification line requires that the classification line not only correctly separates the two categories (training error rate is 0), but also maximizes the classification interval. In the vector space model, the weight of the feature item is often used to comprehensively reflect the contribution of the feature item to the content of the identified text and the ability to distinguish between Japanese text. Since the frequency of each feature item in different Japanese texts meets certain statistical laws, the weight of feature items can be assigned according to the frequency characteristics of the feature items.

Table 1: Test document data processing.

| | Test document 1 | Test document 2 | Test document 3 |
|---|---|---|---|
| P 1 | 0.041 | 2.169 | 0.109 |
| R 1 | 0.043 | 2.133 | 0.14 |
| F 1 | 0.046 | 2.097 | 0.171 |
| P 2 | 0.049 | 2.061 | 0.202 |
| R 2 | 0.052 | 2.025 | 0.233 |
| F 2 | 0.041 | 2.169 | 0.264 |

$$\begin{aligned} &\forall \left| L_i \times L_{k-1} \quad L_{i-1} \times L_{k-1} \quad L_i \times L_k \right| \in C \longrightarrow \\ &\max\left\{ \left| L_i \times L_{k-1} \quad L_{i-1} \times L_{k-1} \quad L_i \times L_k \right| \right\}. \end{aligned} \tag{5}$$

NBC uses the simplest Bayesian network structure. In this model, it is assumed that all attributes $w = 1, 2, \ldots, n$ are conditionally independent of variable $C$, that is, each attribute variable has a class variable as its only parent node. Due to its simple structure, it is sometimes distinguished from a strict Bayesian network and called a naive Bayes classifier. The purpose of the Bayesian classifier is to classify an event and determine whether it belongs to a predetermined category, and the event is expressed as a combination of several feature items. The probability of the event belonging to each category is calculated separately, and the category corresponding to the maximum probability is the category judged by the classifier.

*3.3. Semantic Probability Distribution of Japanese Text.* Corpus term frequency-inverse document frequency (TF-IDF) as a widely used text feature selection method, judges the category of the feature word based on the frequency of the feature word in the Japanese text and the frequency of the feature word in the entire Japanese text data set ability. The category to which a document belongs is only related to the frequency of certain specific words or phrases in the document, and has nothing to do with the position or order of these words or phrases in the document. In other words, if the various semantic units (such as words and phrases) that make up a Japanese text are collectively referred to as "terms," and the frequency of lexical entries in Japanese texts is called "term frequency," then a document implies it. The word frequency information of each term is sufficient to classify it correctly.

$$\left[ f(i) \quad f(i-1) \right] \times \begin{bmatrix} f\left(\left|L_i L_{k-1}\right| + k\right) \\ f\left(\left|L_i L_{k-1}\right| - k\right) \end{bmatrix} - \begin{bmatrix} k & -i \\ i & -k \end{bmatrix} = 0. \tag{6}$$

Knowledge engineering methods mainly rely on linguistic knowledge, and need to compile a large number of inference rules, which is quite complicated to implement. Because of the complexity of natural language, machine understanding of natural language cannot be fully realized at this stage. The current research on Japanese text classification technology mainly focuses on Japanese text classification realized by statistical methods. Compared with knowledge engineering methods, Japanese text classification

based on statistical methods has the characteristics of fast speed and simple implementation, and the accuracy of the classification in Table 2 is also high, which can meet the requirements of general applications.

When the text terminal submits the job to Job Tracker, Task Tracker will use the heartbeat signal to report its status to Job Tracker. If it can perform a new task, then Job Tracker will assign tasks to the Task Tracker that has been prepared. After that, it will communicate with Task Tracker through the return value obtained from the heartbeat signal. Job Tracker will first select tasks for Task Tracker according to the priority list of a certain job. For a job to be executed, it will be divided into several Map tasks, and each Map task will be preallocated to Task Tracker. Map and Reduce have a certain amount of task slots in each Task Tracker node. When the Reduce function allocates tasks, Job Tracker does not think too much about the localization of the data. It just simply selects and executes from the list of tasks to be run.

$$Y(i,j,k) - \left\{ \underset{k+1}{(0,\ldots,0)}, \underset{k+1}{(1,\ldots,1)}, \ldots, \underset{k+1}{(n,\ldots,n)} \right\} = 0, \quad \text{for}\{i, j > k\}. \tag{7}$$

Then calculate the probability distribution of the text topic and the most important feature of the text from the probability distribution of the comment topic. Based on the most important features of the text, the conditional probability distribution is calculated, and then the scoring probability of the text on the existing scoring segment is calculated according to the Bayesian formula. In terms of scoring and testing, the study uses the highest probability score as the text score prediction and uses an experimental design to verify the results. A CSR indicates that when a sentence contains a sequence pattern, it is the conditional probability of comparing sentences, so these CSRs can be used as a classifier.

*3.4. Replacement of Index of Word Formation Mode.* The word-building pattern network has an input layer, an output layer and at least one hidden (middle) layer. The research results show that increasing the number of hidden layers does not necessarily improve the accuracy and expressive ability of the network. The algorithm is a training algorithm for acyclic multi-level networks. Its learning process consists of forward propagation and back propagation. The input value is processed layer by layer from the input layer through the hidden unit after non-linear transformation, and then passed to the output layer. The state of will affect the state of the next layer of network nodes. If the desired output cannot be obtained in the output layer, it will switch to back propagation and modify the weight of each network node to minimize the error signal.

$$\begin{cases} f(i,j \mid i > j) = \left( \dfrac{n-j}{n-k+1}, j \subseteq [k,1] \right), \\ f(i,j \mid i = j) = \underset{k+1}{(0,1,2,\ldots,j-1,} j \subseteq [1,k+1]). \end{cases} \tag{8}$$

The two-dimensional Bayesian model is often widely used in image recognition tasks. The local receptive field of the Bayesian model usually corresponds to a local subregion in the image. There is a weight in this connected local subregion, which can effectively extract some feature attributes in the image, such as the color, directed edges and corners of the graph, etc., to extract these feature attributes. A set of weights is called the convolution kernel of the Bayesian model. The operation of the convolution kernel to move each different area in the image is called convolution. Since the same convolution kernel in Figure 1 has the same weights when processing different areas of the image, Bayesian weights are shared.

It is formed by interconnecting processing units and their undirected signal channels called connections. These processing units have local memory and can complete local operations. Each processing unit has a single output connection. This output can be branched into as many parallel connections as needed, and these parallel connections all output the same signal, that is, the signal of the corresponding processing unit. Naive Bayes has the characteristics of distributed storage of information, global parallelism of operations, and nonlinear processing. It is suitable for learning a complex nonlinear mapping.

$$\iint \frac{n-j}{n-k+1} \times \frac{n}{n+k+1} di\ dj - \iint \frac{n}{n-k+1} / \frac{n+j}{n+k+1} di\ dj = 0. \tag{9}$$

Subject classification refers to automatically classify each Japanese text in a Japanese text collection into a certain category according to a predetermined classification system according to the content of the Japanese text. The text generally uses standard machine learning classifiers, the most commonly used are support vector machines and naive Bayes. In addition, it can also be judged directly by using the more obvious ideographic features, which can be regarded as a rule-based classifier. Sentiment classification at the document level can provide popular opinions about an object, topic or event, but this classification method is difficult to give specific details about what people like or dislike. In spite of this, it still does not prevent it from becoming a popular sentiment analysis method.

## 4. Corpus-Assisted Naive Bayesian Predictive Model Construction of Semantic Word Formation of Japanese Text

*4.1. Language Annotation Corpus Clustering.* The input layer of the language tagging corpus is 0 network nodes, the hidden layer is $h$ network nodes, and the output layer is $m$ network nodes. $n$ is the dimension of the input vector, $m$ is the dimension of the output vector, the number of hidden layer network nodes $h$ can be considered related to the problem, the current research results are still difficult to give the functional relationship between $h$ and the type and scale of the problem. The connection weights between the input layer and the hidden layer, and between the hidden layer and

Table 2: Description of Japanese text classification.

| Conversion operation | Description text | Requirements scores |
|---|---|---|
| Flatmap (function) | Natural language | 44.44 |
| Groupby (function) | Characteristics of fast speed | 45.07 |
| Union (other: RDD[T]) | Knowledge engineering | 45.92 |
| Filter (function) | Statistical methods | 46.73 |
| Mappartitions (function) | Classification technology | 47.37 |
| Sample (withreplacement, seed) | Text classification | 47.74 |
| Groupbykey ([numtasks]) | The accuracy of the classification | 45.43 |

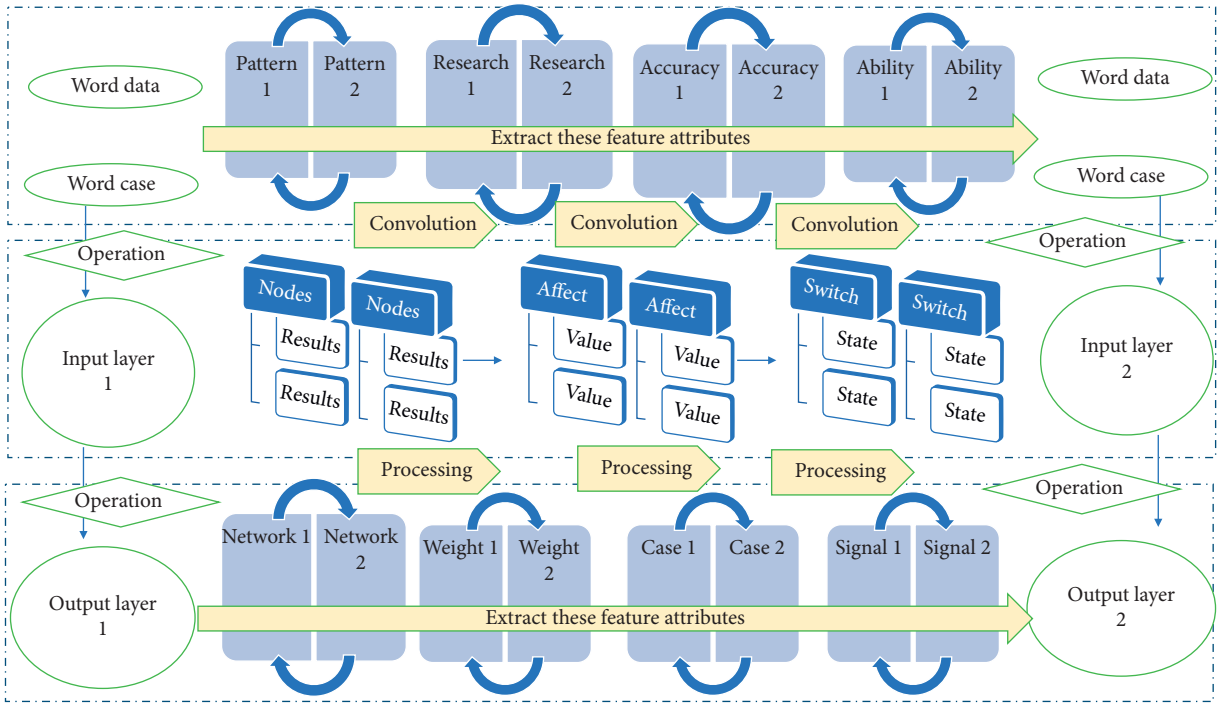


Figure 1: Hierarchical topology of word formation mode.

the output layer are learned from the training samples during the Naive Bayes training stage.

$$\forall q(h-1) \cup q(h+1) = \varnothing, \exists (1-w) \times q(h-1) \overset{x+y>x-y}{+ \cdots +} w \times q(h), h > n\ -1 > 0. \tag{10}$$

The model assumes that each sentence in the review often expresses a certain emotion of the text on a certain feature of the product. In this model, the sentiment of the comment sentence is determined by the sentiment probability distribution, and the sentiment probability distribution can be expressed as a topic model. Such a model contains an assumption that each sentence, that is, a comment, expresses a certain aspect of sentiment, and the generated words are of the same sentiment and theme. The advantage of this model is that the final feature is extracted to the sentence as a unit, so that the preprocessed comment results make it easy to understand the text and carry out corresponding experiments. On the model, each sentence also fully reflects an emotional feature, which is very suitable for processing online text comments.

In the model in Figure 2, the fewer weights, the less data sets can complete the training of the model. At the same time, the fewer weights also indicate that the classification model will have better generalization capabilities. Bayesian has relatively satisfactory translation, scaling and deformation invariance to the input data. The invariance described in the above content allows Bayes to have generalization capabilities that the ordinary network structure does not have, so that it can be successfully used in other fields such as voice and image or Japanese text. The dimension of the feature space is selected by the eigenvalues of the corresponding projected vector.

In more complex network models, Bayes' invariance to input data will be strengthened as the number of layers in the

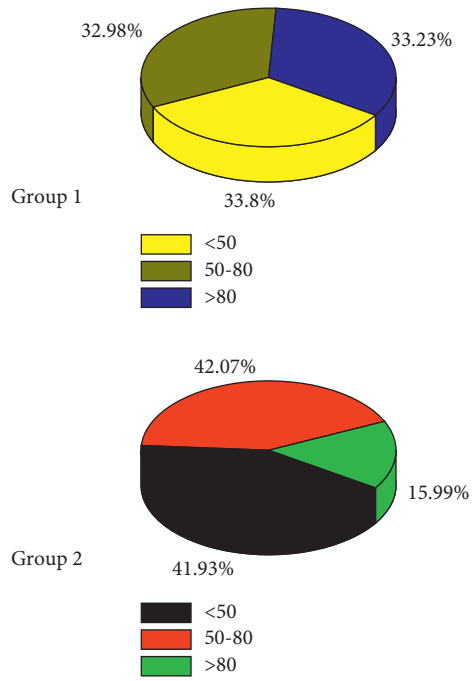


Figure 2: Processing distribution of online text comments.

actual network model increases; in the process of feature extraction on the data set, Bayes has stronger feature extraction ability, Bayesian powerful extraction capabilities can make the network automatically learn and extract the required features. This performance makes the classification model more efficient in practical applications, even if the data set in Table 3 is not processed or simply. The data preprocessing step can be directly applied to the network model.

Sentiment classification is a branch of sentiment orientation recognition. Emotion recognition can be divided into vocabulary-level, sentence-level and text-level emotion recognition. Vocabulary-level sentimentality refers to the recognition of emotions such as emotions, anger, sorrow, and commendation. The sentiment recognition at sentence and chapter level usually judges the polarity of the entire document, and divides the entire Japanese text into praise, derogation, and neutral.

Supervised text-level sentiment classification can be regarded as a special form of topic classification, using various methods of topic classification, including document representation, feature selection, and classification models. However, text-level sentiment classification and topic classification focus on different goals, and the problems to be solved are also different. Topic classification needs to find features that can represent the topic category, while sentiment classification requires semantic analysis of various ways of expressing emotions.

*4.2. Semantic Word Formation Features of Japanese Text.* For the semantic word formation of the obtained Japanese text, it can be processed through the two steps of mapping (Map) and reduction (Reduce). It can be explained in an easy-to-understand method. MapReduce does not have any special ideas in it. It can be regarded as an idea derived from the divide-and-conquer algorithm. When dealing with large tasks, the obtained tasks are first decomposed into it. Several smaller task modules calculate these subtasks separately, and finally summarize the result data obtained from the above calculations.

$$\frac{\sum_{i=1}^{i+k-2} q(i)}{h-1} + \frac{\sum_{i=1}^{i+k-2} q(i-1)}{h-1} + \cdots + \frac{\sum_{i=1}^{i+k-2} q(0)}{h-1} = 1, \quad \text{for}\{h = n\}. \tag{11}$$

It divides a sentence or a document into words and sentence tags. For English, it is not difficult to use spaces to separate words, but there are many other things to consider. For example, when dividing opinion sentences or examples that need to be used, you cannot just divide them by spaces. Part-of-speech tagging and grammatical analysis are techniques used to analyze morphological and syntactic information. Part-of-speech tagging is used to determine the corresponding part-of-speech tag of each word. Similar to word division, part-of-speech tagging is also a sequence labeling problem. Part-of-speech tags, such as adjectives and nouns, are particularly important for opinion mining, because opinion words themselves are adjectives and opinion objects (such as examples or one of their aspects) are nouns or compound nouns.

It is a multilevel classification method, using Figure 3 to transform a complex multi-class classification problem into several simple classification problems to solve. The basic idea is to calculate the feature priority of the features of Japanese text according to a certain function (such as IG), and then sort the priority, and use each feature as the judgment condition (the root node of the subtree) to expand and generate after expansion. The process of classification is to judge according to the conditions of the wild trees. In contrast to this, the grammatical analysis technology obtains syntactic information. Grammatical analysis generates a parse tree, which can express the grammatical structure of a given sentence through the corresponding relations of different components. Compared with part-of-speech tagging, grammatical analysis provides richer structural information. Because the part-of-speech tagging and grammatical analysis in Table 4 have certain similarities and interrelationships, some algorithms have been proposed to handle these two tasks at the same time.

Because Japanese text classification is fundamentally a mapping process, the hallmarks of evaluating the Japanese text classification system are the accuracy of the mapping and the speed of the mapping. The speed of the mapping depends on the complexity of the mapping rules, and the reference for evaluating the accuracy of the mapping is the classification result of the Japanese text after expert thinking and judgment (here it is assumed that the manual classification is completely correct and the factors of personal thinking differences are excluded). This means that the initial 14-dimensional historical experimental data can be represented by a five-dimensional feature space. The closer the results, the higher the accuracy of the classification. Japanese text classification usually uses indicators such as

Table 3: Distribution of data preprocessing steps.

| Distribution | Network model extraction rate | | | | Sentiment orientation recognition error | | | |
|---|---|---|---|---|---|---|---|---|
| Feature 1 | 1.139 | 1.457 | 0.826 | 0.563 | 1.129 | 1.267 | 0.132 | 0.983 |
| Feature 2 | 1.137 | 1.409 | 0.846 | 0.533 | 1.127 | 1.220 | 0.157 | 0.975 |
| Feature 3 | 1.134 | 1.362 | 0.865 | 0.502 | 1.124 | 1.172 | 0.181 | 0.967 |
| Feature 4 | 1.132 | 1.315 | 0.882 | 0.471 | 1.122 | 1.125 | 0.206 | 0.957 |

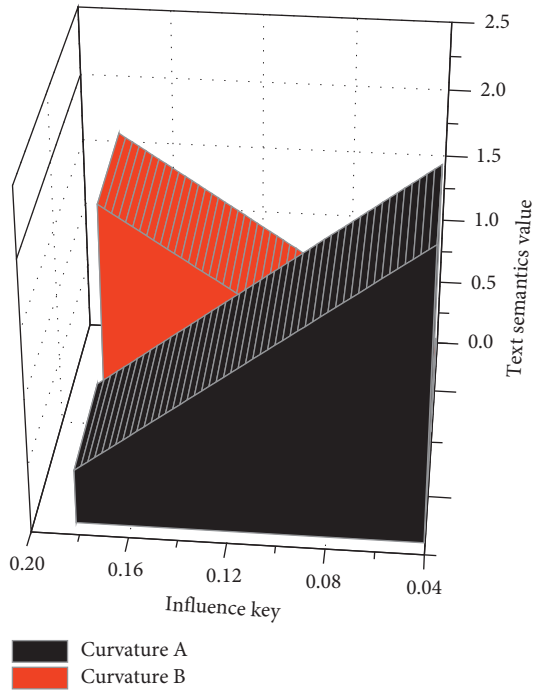


Figure 3: Feature distribution of text semantic word formation.

Table 4: Part-of-speech tagging and syntax analysis algorithms.

| Steps | Analysis algorithms | Code texts |
|---|---|---|
| 1 | Judgment condition $Z'(q(1))$ | Import java.util.Scanner; |
| 2 | Syntactic information | Public class test 06 { |
| 3 | Grammatical analysis $Z(i)$ | Public static void main(String[] args) |
| 4 | The process of $Z'(q(i))$ classification | Scanner input = Scanner (System.in); |
| 5 | The grammatical $Z'(q(i-1))$ structure | Int $a$ = input.nextint(); |
| 6 | Certain similarities and interrelationships | Int $b$ = input.nextint(); |
| 7 | Words and sentence tags $q(i) - q(i-1)$ | Test06 test = new test06(); |
| 8 | Two tasks at the same time $Z(i-1)$ | Int $i$ = test.gongyinshu ($a$, $b$); |
| 9 | Grammatical analysis $\sqrt{a^2 + b^2}$ | System.out.println("$i$" +$i$); |
| 10 | The corresponding part-of-speech $a^2$ | System.out.println ("i"+($a * b/i$)); |
| 11 | Some algorithms $f(q(i)) - f(x)$ | Public class test07 { |
| 12 | A classifier in different $M'(q(i))$ aspects | Public static void main (String[] args) { |
| 13 | Part-of-speech $M'(q(1))$ tagging | Int abccount = 0; |
| 14 | Map function $b^2$ | Int spacecount = 0; |
| 15 | $F$ Value in the field $f(x - x^2 + 1)$ | Int numcount = 0; } |
| 16 | Japanese text classification $M'(q(i))/M'(q(x))$ | If (Character.isletter (ch[i])) { |
| 17 | The manual $x - x^2$ classification | Abccount++; |
| 18 | The $a + b$ accuracy of the classification | }Else if (Character.isdigit (ch[i])) { |

recall (abbreviated as 0), precision (abbreviated as $p$), and $F$ value in the field of information retrieval to evaluate a classifier in different aspects.

*4.3. Naive Bayesian Prediction Bias.* In the Naive Bayesian polynomial model, a document is regarded as a series of ordered collections of words. Assuming that the impact of

article length on a given class is independent, and assuming that any word in the document is independent of its position and context in the text, in this model, the Japanese text vector uses Boolean weights, which are feature items. If it appears in Japanese text, the weight is 1, otherwise it is 0. Suppose the number of features is knives, and treat the Japanese text as an event, which is produced through the experiment of healing, that is, a certain feature appears or does not appear. The Bernoulli model neither considers the order of appearance of feature words, nor does it consider the frequency of feature words in Japanese text.

$$Q(k+i-1)=\begin{cases}\sum_{i=1}^{i+k-2}\sum_{i=1}^{i+k-1} q(i)\cap q(k+i-1)\cap q(k+i-1)\\ \sum_{i=1}^{i+k-1} q(i)\cup q(h-1)\cup q(h+1), \quad i\subseteq[1,n-k+1].\end{cases} \tag{12}$$

Due to the semantic ambiguity of words, it affects the accuracy of the text model to a certain extent. In this chapter, subject terms are used as a supplement to the content of the document, and a text model is established from two different perspectives: keywords and subject terms. Keywords are noun words and phrases that appear in the document and have a significant relationship to the essential meaning of the document, or capture the important characteristics of the document: subject words are not necessary to appear in the document in accordance with the recognized discipline system. Official document classification system or text-customized personalized classification system classifies noun words and phrases in a document or a certain part of a document.

An intuitive view of the 3D features of Fisher's discriminant analysis is shown in the text, most of the data are distributed in three different regional spaces. Through the learning and memory of the sample training set, the naive Bayes classifier can learn the relationship between the category variable of the event and each attribute variable to form the central concept of the training sample, and then use the learned central concept to analyze the unknown category. It can be seen that the Naive Bayes algorithm has good scalability. When faced with the large-scale data problem in Figure 4, it can avoid the process of finding the maximum likelihood and effectively deal with the noise of the data.

The process of establishing the initial interest tree is the process of establishing the model and initializing the model. When the text uses the system for the first time, the system automatically generates an initial user interest tree based on the text registration information and text selection according to the Japanese open directory structure model (assuming the text's initial interest keyword weight is 10). Due to the limited space of the text model, the keywords in the text model should be adjusted to adapt to the changes of text interest and personality over time, so that the word frequency of the most concerned words in the current period of the text remains the highest.

## 5. The Application and Analysis of the Naive Bayes Prediction Model of Japanese Text Semantic Word Formation Assisted by the Corpus

*5.1. Corpus Auxiliary Data Preprocessing.* The result data obtained through the helper function of the corpus will not be directly written to the local disk, it will be stored in the memory buffer of the machine. When the content in the memory buffer exceeds the set threshold, the background thread will start to write the result content to the disk. Before writing the result data to the disk, the data in the memory buffer will be divided into multiple different partitions. These different partitions are determined by which Reduce function the data will eventually be divided into. In all the different partitions, the obtained data will be sorted according to the size of the key value. Assuming there is a Combiner function, it will be aggregated based on the result data obtained after sorting.

$$\left\langle \begin{array}{cc|cc} \sum_{i=1}^{i+k-1} q(i-1) & \sin i & -\sin q(i) & -\sum_{i=1}^{i+k-1}\cos q(i) \\ \cos i & \sum_{i=1}^{i+k-1} q(i) & -\sum_{i=1}^{i+k-1}\sin q(i) & -\cos q(i) \end{array} \right\rangle = \left\langle \begin{array}{cc|cc} \sin i & & -1 & \\ 1 & & \cos i & \end{array} \right\rangle. \tag{13}$$

Combiner plays the role of an optimization function in the entire task. It may not be executed or may be executed multiple times. In the input data, the Combiner function will perform the reduction operation for the same value. In this way, it can reduce data writing can also reduce the consumption of system transmission. The data results in Figure 5 will eventually be integrated into a complete and ordered result data set, which is called for the above process.

The words used in real Japanese texts are often semantically related, such as synonymous relations, synonymous relations, upper and lower relations, and so on. On the other hand, the text's understanding of certain terms may be inconsistent with the author's expression, resulting in different classification topics, thereby affecting the classification results. In order to solve these problems, this paper tries to introduce ontology, using ontology concept hierarchical structure and attribute constraints to match

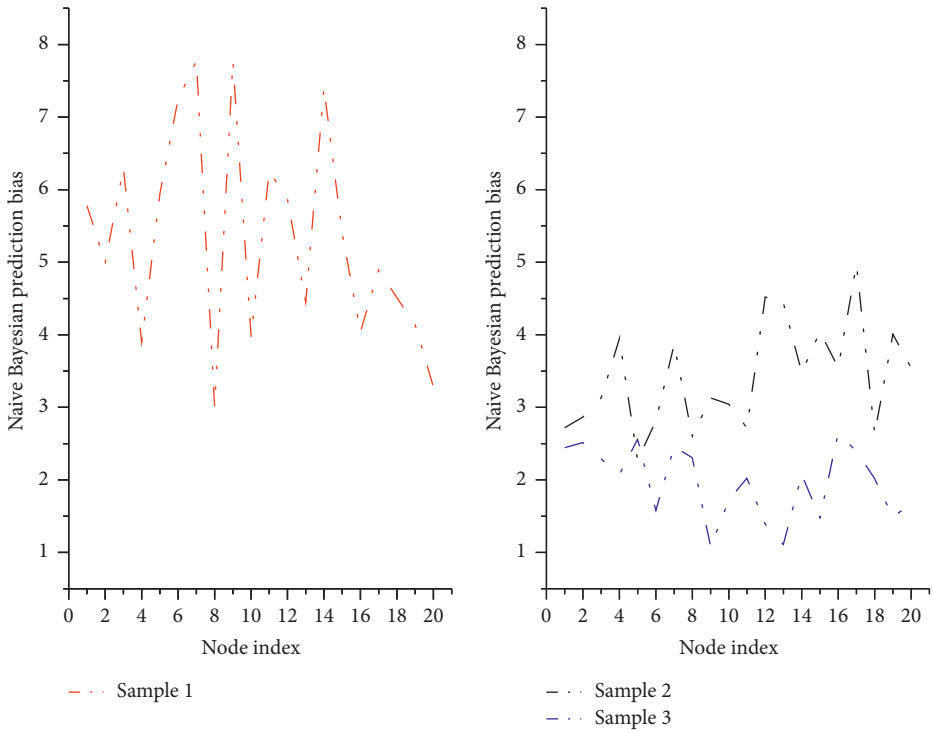


Figure 4: Naive Bayesian prediction deviation distribution.

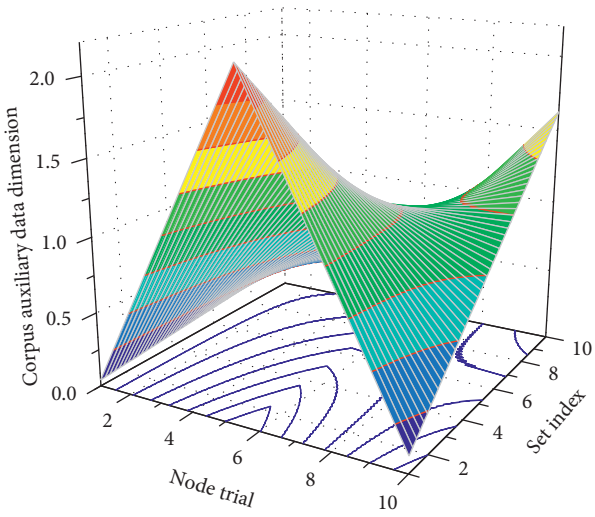


Figure 5: Three-dimensional distribution of corpus auxiliary data.

keywords with domain feature concepts to construct a concept-based vector space model for Japanese text classification.

With the exception of Japanese text comments, other items have constituted natural structured information. In Japanese text reviews, each sentence is also dominated by 1–2 central adjectives or nouns, with very colloquial structure and different forms. The part-of-speech tagging and grammatical analysis of such Japanese texts must be relatively inefficient. Using the Japanese word segmentation model with stop words will achieve good results.

$$\begin{cases} h - [k \times q] \times w = 0, \\ q - \dfrac{k+1}{n-k+2} = 0 \quad \text{if} \quad \begin{matrix} k, q \longrightarrow 1, 1 \\ w, d \longrightarrow 0, 0 \end{matrix}, \\ w - k \times d - n = 0, \end{cases} \tag{14}$$

We predict the text sample data $X$ of each location class label given in the sample. Naive Bayes formula can be used to predict the class with the largest posterior probability to

which text $X$ belongs. The conditional probability of the category is compared with the product result of the posterior probability obtained through the training set to determine the text life cycle stage of the text. It can be seen that if a text attribute or text category is added, in order to ensure that the text attribute is completely "learned" by the naive Bayes algorithm, it is necessary to have a larger number of samples in the training set as a basis. There is a probability that the sample information statistics is zero, which leads to a large deviation in the prediction results; in addition, the increased number of text attributes or categories will cause the complexity of the "learning" process of the Naive Bayes algorithm to be geometrical Increase, which requires accurate text classification with the help of a computer.

*5.2. Naive Bayes Prediction Simulation Realization.* The original corpus of the experiment is book reviews downloaded from the Internet, with a total of 51 58 reviews, which are manually divided into two types: positive reviews and negative reviews. Among them, positive reviews, that is, artificially judged as praise, happy, or implicitly praised a total of 2,600, and negative reviews, that is, artificially judged as depreciative, angry, or implicitly derogatory, a total of 2558. The word segmentation adopts the developed system.

After word segmentation, regardless of punctuation, there are a total of 11744 words (characters). The average number of words (characters) per Japanese text is 21.6. As some comments are too short, after preprocessing, 5089 pieces of corpus are obtained, including 2,576 pieces of positive evaluation corpus and 25 13 pieces of negative evaluation corpus. Each type of Japanese text is randomly divided into four equally, three of which are used as the training set and one is used as the test set. Experiments show that the stop vocabulary table can reduce the dimensionality of the feature space, and will have a positive effect on improving the classification accuracy of the classifier.

When using adjectives, adverbs, nouns, and verbs as candidate features, the newly added semantic features can effectively improve the emotion recognition rate under each feature selection method. Especially when there are 200–400 features with a small number of features, the recognition rate of Figure 6 improves quickly, indicating that most of the newly added semantic features rank high in the feature table, and can effectively improve the emotion recognition rate. If the data is projected into a five-dimensional feature space, the different health patterns can become more discrete.

By taking the average value of the element sum of the currentarray, it is found that in the bag-of-words model, the average word frequency of the Japanese text review document is 8, which is equivalent to the word segmentation result of a single sentence. This leads us to believe that the topic relevance in the comment data sentence is relatively small. It can be considered that the topic generation of a single sentence is similar to the topic generation of the selected words one by one. Therefore, the LDA model can be used to replace the sLDA model for topic modeling.

$$\frac{\prod_{q(i)-q(i-1)} Z(i-1) + \cdots + Z(i)}{Z'(q(i))} - \frac{\prod_{q(i)-q(i-1)} M'(q(i)) \times M''(q(i))}{Z''(q(i))} = 0. \tag{15}$$

Information gain is a relatively common feature selection method in the process of data preprocessing. The feature with larger information gain value is selected as the feature subset. When performing data preprocessing on the Japanese text training set, information gain is a feature selection method that only considers how much information the feature words bring to the whole world, and does not consider what kind of information changes the feature words bring to a specific category.

When processing an unbalanced Japanese text data set, because the features of each category are unevenly distributed in the data set, the feature subset selected by the traditional information gain method is unreasonable, which affects the subsequent classification results. Aiming at the shortcomings of information gain, this paper adds a word frequency adjustment factor based on the word frequency distribution of feature words in the data set, and redefines the conditional entropy.

$$\frac{M'(q(1)) \times M''(q(1))}{M'(q(i))} + \frac{M'(q(2)) \times M''(q(2))}{M'(q(i))}, \cdots + \frac{M'(q(i)) \times M''(q(i))}{M'(q(i))} - 1 = 0''. \tag{16}$$

It shows that when using all words as candidate features, semantic features can also effectively improve the recognition rate. In particular, it is noted that the use of mutual information (MI) for feature selection has more obvious changes. When 200 features are used, semantic features are used. The recognition rate of feature MI has reached more than 80%, which is nearly 5% higher than the MI method without semantic features, and at 90.1 000 features, the MI method with semantic features surpasses all cases. Under the recognition rate. The reason for the obvious change of the MI method is that it has a preference for low-frequency words with strong classification ability, and many of the semantic features are low-frequency features, so that the features used by the classifier are mostly generated semantic features, and these features perform well.

*5.3. Case Application and Analysis.* The results obtained by using this model on the validation set are shown in the text. It can be seen that the Japanese text classification based on the hybrid naive Bayes model proposed in this paper has also obtained good results on the validation set. In order to further improve the weight value of topic words, the TF-IDF

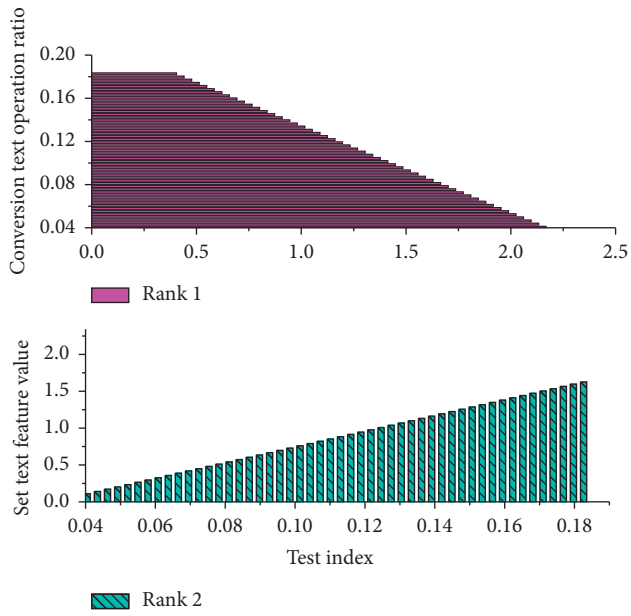


Figure 6: Dimensional distribution of Japanese text feature space.

value is introduced in the generation of the word embedding layer. The TF-IDF value can measure the particularity of words for a type of document and emphasize the high-frequency feature words in the category. And it is easier to divide, and this distinguishing feature will be used in the subsequent diagnosis process.

The word vector generated by word2vec and the TF-IDF value are weighted to form a word embedding layer. In order to improve the accuracy of Japanese text classification, combining the feature extraction performance of the two models, not only can extract the local features of the Japanese text, but also capture the contextual semantic information. At the same time, the dropout random inactivation strategy is added to improve the model's resistance to over-simulation.

$$\begin{vmatrix} Z'(q(i)) & \sqrt{a^2+b^2} & 0 \\ \sqrt{a^2+b^2} & Z'(q(i-1)) & \sqrt{a^2+b^2} \\ 0 & \sqrt{a^2+b^2} & Z'(q(1)) \end{vmatrix} \overset{f(qi)-f(x)=0}{\rightleftarrows} \begin{vmatrix} f\left(x-x^2\right) & a+b & 1 \\ a+b & f\left(x-x^2-1\right) & a+b \\ 1 & a+b & f\left(x-x^2+1\right) \end{vmatrix}. \tag{17}$$

For the web Japanese text data set on the Internet, the improved information gain feature selection method proposed in this paper is used to perform data preprocessing on the Japanese text data set, and the dimensionality of the data set is reduced, and the strong distinguishing and representative ones are selected. Then, using the ant colony optimization algorithm to iteratively optimize the weights of the weighted naive Bayes classification algorithm, find the global optimal solution of the weights of the weighted naive Bayes classification model, and balance the correlation between the attributes Negative impact on weights; finally, the combination constitutes the optimized weighted naive Bayes classification model proposed in this chapter.

We can add the custom dictionary needed in the Japanese text comment research through the two operations in Figure 7. The first is to add the entire custom dictionary document, the format is the same as the commonly used stop vocabulary, the difference is that the latter is under research. The noise words that need to be eliminated, and the words in the custom dictionary are to treat these words as a whole in the precise word segmentation and do not separate them.

The second is to temporarily add a piece of identification information to the dictionary, and treat this information as a whole word, which is more suitable for agile algorithm adjustment. Through the analysis of historical data, the general laws implied in the data are extracted and used to predict the nature and types of future data. There are similarities between prediction and classification. Both use historical data to speculate on unknown data. The biggest difference between the two is the output value. The classification output is discrete data, while the regression prediction output is continuous.

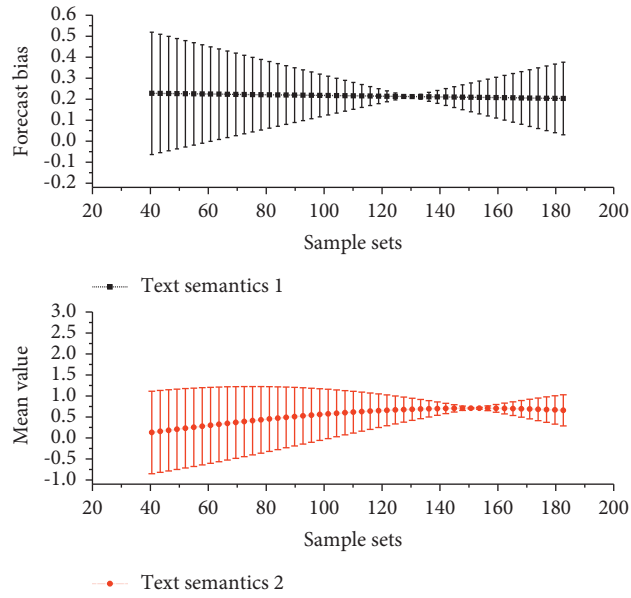


Figure 7: The text classification statistical distribution of the naive Bayes model.

$$\lim_{x,i\longrightarrow\infty}\frac{M'(q(1))+M'(q(2))+M'(q(i))+\cdots+M''(q(x))}{\prod_{q(i)-q(i-1)}M'(q(i))+M''(q(i))}-\lim_{x\longrightarrow\infty}\frac{M'(q(i))}{M'(q(x))}=0. \tag{18}$$

The influence of the number of iterations, in theory, the higher the number of iterations can bring me higher reliability, but through experiments, this research found that the accuracy rate will decrease after the number of iterations exceeds a certain amount. The clustering process is the process of regrouping the original data, so that after regrouping, the data between groups are obviously different, and the data in each group are as similar as possible. The features and attributes recorded by the existing data are used as the training set for data classification, the classification model is established through a supervised "learning" process, and the data labeled with features and attributes are classified to make inferences. It can also be understood that the data for clustering has no prerequisites, and the application of classification requires the data to have certain prerequisites.

## 6. Conclusion

Aiming at the problem that the original vocabulary features cannot fully adapt to the classification, this article starts from the emotional expression of Japanese texts, and proposes to use semantic features to supplement the description of Japanese texts. By adding semantic features to the Japanese texts for emotional description, the extracted features are more conducive to emotion recognition. Japanese text classification has a very close relationship with information retrieval. It borrows many retrieval methods and techniques to promote the development of classification. This paper analyzes the text interest model, establishes the text interest model from the perspective of key words and topic words, and proposes an adjustment algorithm of the text interest model to make the classification result more in line with the text intent. Based on the ontology to obtain the concept features, the concept space is used to replace the keyword space, introduce ontology concepts, modeling primitives, construction methods and construction tools, and use ontology construction tool to establish ontology in the education field. Analyzing the meaning of ontology applied to Japanese text classification, that is, solving the problem of terminology confusion. At the same time, in view of the problem that traditional Bayesian classifiers need to perform repeated searches on features in the two stages of feature selection and training classifiers, which is not conducive to the problem of data acquisition by the system, a statistical corpus module is designed, which can obtain features in one time. The information needed in the entire classification process simplifies the search process, the entire naive Bayes classification process is designed and implemented, and a naive Bayes classification platform is completed for emotion recognition. The experimental results show that under different stop vocabulary lists and different feature selection methods, the new semantic features proposed in this paper can effectively improve the text recognition rate.

## Data Availability

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

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] K. Matsumoto, S. Tsuchiya, and T. Miyake, "Flame prediction based on harmful expression judgement using distributed representation," *International Journal of Technology and Engineering Studies*, vol. 4, no. 1, pp. 7–15, 2018.

[2] D. Gerz, I. Vulić, E. Ponti, J. Naradowsky, R. Reichart, and A. Korhonen, "Language modeling for morphologically rich languages: character-aware modeling for word-level prediction," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 451–465, 2018.

[3] A. S. Dylman and M. Kikutani, "The role of semantic processing in reading Japanese orthographies: an investigation using a script-switch paradigm," *Reading and Writing*, vol. 31, no. 3, pp. 503–531, 2018.

[4] M. Al-Smadi, M. Al-Ayyoub, Y. Jararweh, and O. Qawasmeh, "Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features," *Information Processing & Management*, vol. 56, no. 2, pp. 308–319, 2019.

[5] E. Matsuki, Y. Hino, and D. Jared, "Understanding semantic accents in Japanese-English bilinguals: A feature-based approach," *Bilingualism: Language and Cognition*, vol. 24, no. 1, pp. 137–153, 2021.

[6] A. Fronzetti Colladon, C. A. D'Angelo, and P. A. Gloor, "Predicting the future success of scientific publications through social network and semantic analysis," *Scientometrics*, vol. 124, no. 1, pp. 357–377, 2020.

[7] L. K. Branting, C. Pfeifer, B. Brown et al., "Scalable and explainable legal prediction," *Artificial Intelligence and Law*, vol. 29, no. 2, pp. 213–238, 2021.

[8] S. Tibi, A. A. Edwards, C. Schatschneider, and J. R. Kirby, "Predicting Arabic word reading: A cross-classified generalized random-effects analysis showing the critical role of morphology," *Annals of Dyslexia*, vol. 70, no. 2, pp. 200–219, 2020.

[9] S. R. Jammalamadaka, J. Qiu, and N. Ning, "Predicting a stock portfolio with the multivariate Bayesian structural time series model: do news or emotions matter?" *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 81–104, 2019.

[10] Z. Zeng, Y. Deng, and X. Li, "Natural language processing for EHR-based computational phenotyping," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 139–153, 2018.

[11] J. Chambua, Z. Niu, A. Yousif, and J. Mbelwa, "Tensor factorization method based on review text semantic similarity for rating prediction," *Expert Systems with Applications*, vol. 114, pp. 629–638, 2018.

[12] E. Seok Nee, C. Sern Choong, A. Shahrizan Abdul Ghani, A. P. P. Abdul Majeed, A. Adam, and M. Furqan, "Text-based emotion prediction system using machine learning approach," *IOP Conference Series: Materials Science and Engineering*, vol. 769, no. 1, p. 012022, 2020.

[13] A. Krishna, B. Santra, A. Gupta, P. Satuluri, and P. Goyal, "A graph-based framework for structured prediction tasks in Sanskrit," *Computational Linguistics*, vol. 46, no. 4, pp. 785–845, 2021.

[14] H. Kimura and H. Narita, "Compound w-questions and fragment answers in Japanese: implications for the nature of ellipsis," *Linguistic Inquiry*, vol. 52, no. 1, pp. 195–209, 2021.

[15] Y. Sato and M. Maeda, "Syntactic head movement in Japanese: evidence from verb-echo answers and negative scope reversal," *Linguistic Inquiry*, vol. 52, no. 2, pp. 359–376, 2021.

[16] N. Kwon and S. Yu, "Experimental evidence for the productivity of total reduplication in Japanese ideophones and ordinary vocabulary," *Language Sciences*, vol. 66, pp. 166–182, 2018.

[17] K. Akita, "A typology of depiction marking," *Studies in Language*, vol. 45, no. 4, pp. 865–886, 2021.

[18] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3767–3780, 2019.

[19] Y. Zhang, J. Fu, and D. She, "Text emotion distribution learning via multi-task convolutional neural network," *IJCAI*, pp. 4595–4601, 2018.

[20] J. Horvat, "From Paul the Octopus to achilles the cat-proper names of animals which predict the outcomes of sports competitions," *Folia Onomastica Croatica*, vol. 29, no. 29, pp. 73–121, 2020.

[21] X. Zhuo, F. Fraundorfer, F. Kurz, and P. Reinartz, "Automatic Annotation of airborne images by label propagation based on a bayesian-CRF model," *Remote Sensing*, vol. 11, no. 2, p. 145, 2019.

[22] Y. H. Chen, F. Zhang, and W. L. Zuo, "Deep image annotation and classification by fusing multi-modal semantic topics[J]," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 1, pp. 392–412, 2018.

[23] I. R. Braun and C. J. Lawrence-Dill, "Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction," *Frontiers of Plant Science*, vol. 10, p. 1629, 2020.

[24] A. Joshi, S. Karimi, and R. Sparks, "Survey of text-based epidemic intelligence: A computational linguistics perspective," *ACM Computing Surveys*, vol. 52, no. 6, pp. 16–19, 2019.

[25] B. Thompson, S. G. Roberts, and G. Lupyan, "Cultural influences on word meanings revealed through large-scale semantic alignment," *Nature Human Behaviour*, vol. 4, no. 10, pp. 1029–1038, 2020.