International Journal of Web-Based Learning and Teaching Technologies
Volume 19 • Issue 1

Feature Extraction of Dialogue Text Based on Big Data and Machine Learning

Xuelin Liu, Weifang University, China
Hua Zhang, Weifang University, China*
Yue Cheng, Adamson University, Philippines

ABSTRACT

In this article, a dialogue text feature extraction model based on big data and machine learning is constructed, which transforms the high-dimensional space of text features into the low-dimensional space that is easy to process, so that the best feature words can be selected to represent the document set. Tests show that in most cases, the classification accuracy of this model is higher than 88%, and the recall rate is higher than 85%, thus achieving the goal of higher classification accuracy with less computation. When extracting the features of dialogue texts, there is no need for preprocessing, just count the data such as lexical composition, sentence length and sentence-to-sentence relationship of the target text, and make linear analysis to obtain key indicators and weights. Based on this, the classification model can achieve good results, thus effectively reducing the workload and computation of text classification.

KEYWORDS

Big data, Dialogue text, Feature extraction, Machine learning, Text features

INTRODUCTION

With the continuous development of computer and internet technology, information and data are exploding. How to effectively use these huge and disorderly collections of information, classify it accurately and efficiently, and summarize valuable information derived from it has become an urgent problem (Kang & Youn, 2020). In response, research into text classification technology has come into being. Text classification is the sorting of each piece of text into a predefined category, so that users can quickly and conveniently obtain the required information by some means of searching while reading the text. (Barbantan et al., 2016). Today's daily information exchange often hides some clues or evidence, and we often lack the means of extracting useful knowledge from this information often lacks means (Bharti & Singh, 2015). For example, content collected from social media can be used to extract information about a person's interests, behavior, and habits; information about the internal communication of a company can be extracted from its e-mail; information about user behavior

DOI: 10.4018/IJWLTT.337602 *Corresponding Author

can be extracted from a website's network traffic. These text information analyses are based on text mining, text analysis technology, and classification technology in machine learning. Text mining is a technology for extracting useful information from text. It can be used to find patterns and rules in text. Text analysis technology, which is also designed for obtaining information from text, can be used to analyze the semantics and syntax of text. Classification technology in machine learning is a technology for extract useful information from data. It discerns patterns and rules in data, so as to extract useful information.

In both document-based processing and text processing, it is necessary to analyze the text itself (Karthikeyan, 2019). The information content and format documents are diverse and complex. Many data in the database are structured, such as relational data and data warehouse data, are structured (Lee, 2019). But some of the data in the document are semi-structured, and more are unstructured. Unstructured data can't be processed directly, so text data must be converted into data form that can be recognized by the computer by using the established corresponding data model (Oskouie, 2014). The quantity and form of online data have changed correspondingly; while people are faced with these rich resources, there is a contradiction between people's ever-growing demand for information and the increasing difficulty of access to the required information (Jiang et al., 2022). It is a great challenge in the field of information science to organize and mine these data effectively in order to find the information users need more quickly, accurately, and comprehensively. In this paper, a dialogue text feature extraction model based on big data and machine learning is constructed, which transforms the high-dimensional space of text features into a low-dimensional space that is easy to process, so that the best feature words can be selected to represent the document set. In high-dimensional space, the distribution of data is relatively complex, and the computational and storage costs will also increase accordingly. On the contrary, in low dimensional spaces, the number of feature vectors is relatively small. At the same time, in low dimensional space, we are more likely to discover hidden relationships and information in text data, and low-dimensional space, is also easier to visualize and intuitively display.

The unique feature of dialogue texts is that they usually contain several aspects of information, such as the personality traits, emotional states, language usage habits, and established language rules of the dialogue participants. This information needs to be extracted and represented through special methods in order to better understand and analyze the dialogue text. In addition, dialogue text features generally involve multiple text segments. These fragments may belong to different speakers, so it is necessary to consider how to effectively integrate and utilize this information to obtain a more comprehensive and accurate description of the conversation as a whole. Because the data set of dialogue text is very large, traditional text processing methods may encounter the problem of "the curse of dimensionality" in the course of processing. Therefore, it is very important to use methods such as machine learning for feature extraction and comprehensive analysis of dialogue texts (Lv et al., 2018). Feature extraction mainly includes the assumption that each feature word is independent of each other and process each feature word in the original feature set separately (Garla & Brandt, 2012). According to the calculation of feature selection function, each feature word is given a different weight, and the best feature word is selected from it (Calvo, 2018). Although there are various forms of online data, text data mining is very important, because the main carrier of data is text, and other forms of data can be converted into text (Wang, 2021). This paper constructs a feature extraction model for dialogue text based on big data and machine learning. It mainly includes the following innovations and contributions:

(1) According to the attributes of characters in documents in linguistics, this model uses language rules, fully considers a large number of synonymy and polysemy phenomena in language, and some key factors such as commendatory tendency, so as to improve the accuracy of feature extraction and text filtering.

(2) The model transforms the high-dimensional space of text features into a low-dimensional space which is easy to handle, so that the best feature words can be selected to represent the document set.

RELATED WORK

Common feature word selection methods include the text frequency-based selection method, the information gain-based selection method, the mutual information-based selection method, and the statistics-based selection method. Of course, these methods also have some shortcomings. This processing is based on the independence of each feature word, but in fact, the feature words are related, so the method of ignoring the relationship between feature words will sometimes affect the dimension-reduction effect.

Ma et al. (2020) put forward an improved TF-IDF(Term Frequency-Inverse Document Frequency) method combining information entropy. This method adjusts the weight calculation of TF-IDF feature items by combining the information distribution entropy of feature items between classes and within classes; thus this method avoids the defect of giving larger weights to those feature items that do not contribute to classification, and can calculate the weight of text feature items more effectively. However, this method can't effectively reflect the importance of words or phrases and the distribution of characteristic values. Zhou et al. (2019) put forward an expected cross entropy algorithm based on inter-class concentration and intra-class dispersion, which organically combines the uniformity of feature items between classes and within classes and produces good results in text classification, but it needs the support of high-dimensional feature subsets. Qian et al. (2011) proposed a text classification algorithm based on support vector machine according to its advantages in solving small-sample, nonlinear and high-dimensional pattern recognition problems. Lin et al. (2011) put forward an automatic text classification algorithm based on sequence, which makes use of the correlation between sub-pattern concept nodes in the text to generate feature sequences and then classify the text, thus improving the classification effect. Chambua et al. (2018) added word frequency information, document frequency information, and a category correlation factor to improve the performance of the mutual information method in feature weighting, and proposed an improved feature weighting method based on mutual information. This method has better classification performance than the traditional feature weighting method. Touzani et al. (2021) proposed constructing adjacent weighted graphs and their complement graphs on training data sets by using category information so that the projections of sample points belonging to the same category are as close as possible, and the projections of sample points not belonging to the same category are as far as possible. This method not only can obtain the global structure information of the text space, but can also retain the local structure information; however, it has high computational complexity. Chapman et al. (2012) modified the expression to increase the weight of frequently appearing terms and used the improved method to extract features and applied them to the classification algorithm, which showed a good classification effect. Xie et al. (2015) proposed a method to qualitatively evaluate the performance of the feature selection function and defined a set of basic constraints related to classification information. Lee et al. (2012) proposed a hybrid text feature dimensionality reduction method based on feature selection and feature extraction, which used a new feature extraction method based on extracting the feature set twice, and realized the effective dimensionality reduction of text features on the premise of minimizing information loss.

For text classification, the dimensionality of candidate feature sets after preprocessing is significant for classification algorithms, which has led to research on feature dimensionality reduction. General feature dimensionality reduction is achieved through feature extraction, with the main purpose of removing meaningless features from the feature set for text classification, in order to improve classification performance. Traditional text feature extraction methods are based on mathematical statistics, neglecting the semantic relationships between words in the text. This article attempts to combine semantic information with traditional feature extraction methods to achieve the

goal of integrating mathematical statistical information and semantic information into classification algorithms and thus improve the effectiveness of text classification.

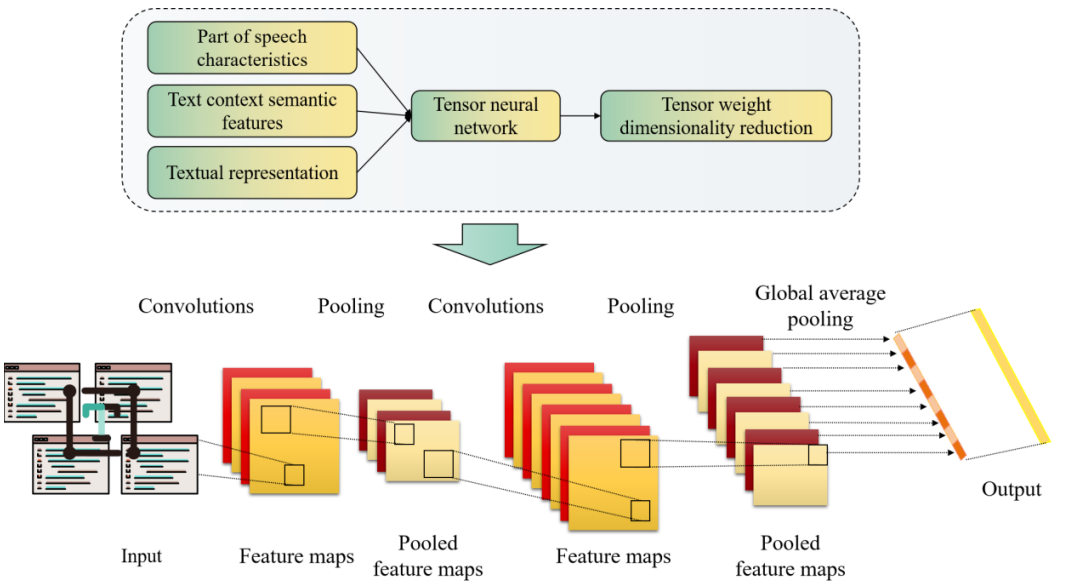
METHODOLOGY

Data Mining of Text Classification Based on Machine Learning

Text classification is a technology based on the framework of natural language processing technology to judge the text category labels such as words, sentences, paragraphs, etc. It belongs to the supervised learning method. The text comes from a custom category, and the category labels can be any number. According to different categories, text classification includes single-level classification and multi-level classification, with one category label for single-level classification and multiple categories labels for multilevel classification (Chakroborty & Saha, 2010). The word segmentation method based on string matching is most widely used in practical systems; it is also called the *mechanical word segmentation method*. Its main aim is to build a dictionary large enough to match the string to be analyzed with the entries in the dictionary according to a certain algorithm. If it is found, a word will be divided if the matching is successful.

For the process of text classification, a classifier will perform text preprocessing in the training set during the training phase. The process of feature extraction and document representation is established by some classification algorithm. In the classification phase, we also process the test set in the above process, and then determine the category of the document set according to the classifier. *Mutual information* is a statistical indicator used to measure the correlation between two variables. It can be used to find the correlation between features, thus helping us better understand the data. From the perspective of information theory, mutual information measures the amount of information brought by whether words are included in categories or not. When the term is the best feature to determine the category, the mutual information reaches the maximum value. And when only a certain document belongs to that current category, word items appear in the document. The process of word vector feature extraction based on machine learning is shown in Figure 1.

Figure 1. The process of word vector feature extraction based on machine learning



Each format of multimedia data needs specific retrieval methods, including specific underlying features, similarity methods, and query patterns; therefore it requires retrieval tools to support as many data types as possible. Different kinds of multimedia data have different structural characteristics, so special retrieval methods are needed to index and retrieve the information. For example, image data can be retrieved by using some image processing techniques, such as RGB values or gray values; video data can be searched using features such as motion direction, geometric texture, or color analysis; audio data can be retrieved using frequency characteristics, sound pressure characteristics, or time distribution characteristics. In other words, each format of multimedia data needs a specific retrieval method to effectively retrieve. Different types of retrieval have different basic contents and different problems to be solved. In the offline stage, multimedia data preprocessing, media content analysis, data indexing, and so on need to be completed, laying the foundation for efficient and accurate online retrieval. Text retrieval is the process of searching out the texts that meet the user's query conditions from a large number of relatively stable text sources according to the specific query put forward by the user and sorting them according to the degree to which they satisfy the query (Wang et al., 2016). In the online retrieval stage, according to the query submitted by the user, the search engine finds out the media data related to the query from the index and uses the sorting technology to generate the result sequence. Introducing user feedback in the retrieval process is an effective way to improve retrieval accuracy. The retrieval system supports users in continuing to participate in the retrieval process after inputting the query and in marking the samples as related or unrelated to the current retrieval results to clarify their information needs. Then the system improves the retrieval model, adjusts the retrieval strategy, and updates the retrieval results according to the feedback from users.

The high-dimensional feature vector problem in the vector space model is an important problem to be solved in text classification. The number of preprocessed initial feature items is generally vast, and many of the items are useless for text classification. This unmanageable volume not only increases the running time of the classification algorithm but also affects the usefulness of text classification. Therefore, many researchers are seeking ways to effectively reduce the number of feature words—that is, to effectively extract features. In the vector space model, the similarity between texts can be expressed by some distance between vectors. It is usually calculated by the inner product or cosine of the included angle between vectors, as shown in formula (1):

$$D_1 = \left(w_{11}, w_{12}, w_{13}, \ldots, w_{1n}\right) \qquad D_2 = \left(w_{21}, w_{22}, w_{23}, \ldots, w_{2n}\right) \tag{1}$$

The inner product formula is shown in Formula (2):

$$Sim\left(D_1, D_2\right) = \sum_{k=1}^{n} w_{1k} w_{2k} \tag{2}$$

The included angle cosine formula is shown in Formula (3):

$$Sim\left(D_1, D_2\right) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}} \tag{3}$$

The similarity between the query and the text is calculated in the same way. Words that often appear in the same window are likely to form a phrase, and this phrase can better express the main

point of a piece of text. For quadrilaterals, four words appear in the same window in pairs, indicating that the scope of their appearance occupies a larger proportion of this paper and expresses more semantics of context.

The task of text classification based on machine learning is essentially a process of function mapping. Feature extraction is the process of selecting some subsets from the pre-processed words in the training set, and these subsets will be used as features in the subsequent text classification process. In the text classification task, we can reduce the term space and retain more effective terms by removing the stop words, stem extraction, and part-of-speech tagging and using other preprocessing technologies to improve the efficiency of classifier training and application (Dong et al., 2018). Moreover, preprocessing can remove noise features and then improve the accuracy of text classification. *Noise features* refers to those features that will reduce the accuracy of text classification after being added. Generally speaking, a noise feature can be distinguished by calculating the information gain or information gain ratio of the feature. If the information gain or information gain ratio of a feature is relatively low, it can be considered a noise feature and can be removed. Before text classification, it is necessary to preprocess the original data, eliminate the noise in the original data, and transform it into a standard format. In text preprocessing, word segmentation algorithm based on dictionary matching or statistics can be used for text analysis. The word segmentation algorithm based on dictionary matching or statistics can effectively and accurately separate each word in the text and is beneficial for breaking down the text into meaningful phrases that can be captured semantically. It can also help eliminate some errors in text, such as spelling errors and punctuation errors, so as to improve the accuracy of text processing.

When extracting word vector features, we can infer the meaning of a word through the context and further judge the emotional content of the word. As for the word similarity calculation of the statistical model, since the statistical model uses word vector to assign each word a multidimensional vector to represent its position in the lexicon, the similarity between two words can be quantified by calculating the angle between two-word vectors:

$$sim_1\left(W_1, W_2\right) = \cos\theta = \frac{\vec{v}_1 * \vec{v}_2}{\left|\vec{v}_1\right| * \left|\vec{v}_2\right|} \tag{4}$$

where $\vec{v}_1$ and $\vec{v}_2$ are word vectors of W_1 and W_2. When two words are synonyms or identical words, the angle between the two-word vectors is 0 and the word similarity is 1. When the two words are completely different, the angle between the two words vectors is 90 and the similarity of words is close to 0. Every sememe forms a hierarchical system according to a hyponymic structure. *Semantic element* refers to the smallest semantic unit in semantic analysis. A semantic element can be a word, a phrase, or even a sentence. The sememe can form a hierarchical system according to the upper and lower structures, so as to better express the semantics. In this system, the distance between words reflects the semantic similarity between words, as shown in formula (5):

$$\begin{cases} \lim\limits_{d\to\infty} sim\left(W_1, W_2\right) = 0 \\ \lim\limits_{d\to 0} sim\left(W_1, W_2\right) = 1 \end{cases} \tag{5}$$

The shorter the distance between words, the more similar they are, and the longer the distance between words, the less similar they are. Word similarity can be quantified by calculating semantic distance. The calculation of the semantic distance between two sememes is shown in formula (6):

$$sim_2\left(W_1, W_2\right) = \frac{\alpha}{d\left(W_1, W_2\right) + \alpha} \tag{6}$$

where W_1 and W_2 represent two sememes, and $d\left(W_1, W_2\right)$ represents the length of the shortest path of W_2 and W_2 in sememes hierarchy. α represents an adjustable parameter. With the process of simultaneous transmission of past and future information, we can effectively utilize the past features—that is, the forward transmission state and the future features, or the backward transmission state, and get all the required hidden state values through the expanded forward and backward networks in a convenient period of time.

The most important feature of the vector model is that it can easily calculate the similarity between any two vectors—that is, the similarity between the texts corresponding to the vectors. In information retrieval, if two vectors are similar, their corresponding texts are semantically related. When all texts and queries are expressed in vector form, it is a common method in modern information retrieval systems to compare the similarity of a particular query vector with all text vectors and sort the texts in descending order according to the similarity (Figueiredo et al., 2011). In the process of indexing, first, a feature extraction module is used to extract text features. Next, the automatic text recognition engine with targeted training will recognize the text data set, and then the indexing engine will build a text index for the recognized output and store the text index.

In text representation, it is unnecessary to consider the relationship between the structural elements of the text; each element is regarded as an independent item, and the calculation weight is set according to the frequency of words. It is necessary to reduce the dimension of the high-dimensional feature matrix in space so as to lessen the vastness of the original data, which would otherwise overload the computing power of the computer (Vicient et al., 2013). The vector model regards the text space as a vector space composed of orthogonal entry vectors. Obviously, the dimensions of vectors are often very large, and the features are not prominent, so we hope to find such a feature subset without affecting the accuracy of classification. This set is composed of some terms that can be used to effectively distinguish categories, and these terms are selected from the feature word set, so as to reduce the vector dimension, simplify the calculation, and improve the accuracy and efficiency of text classification.

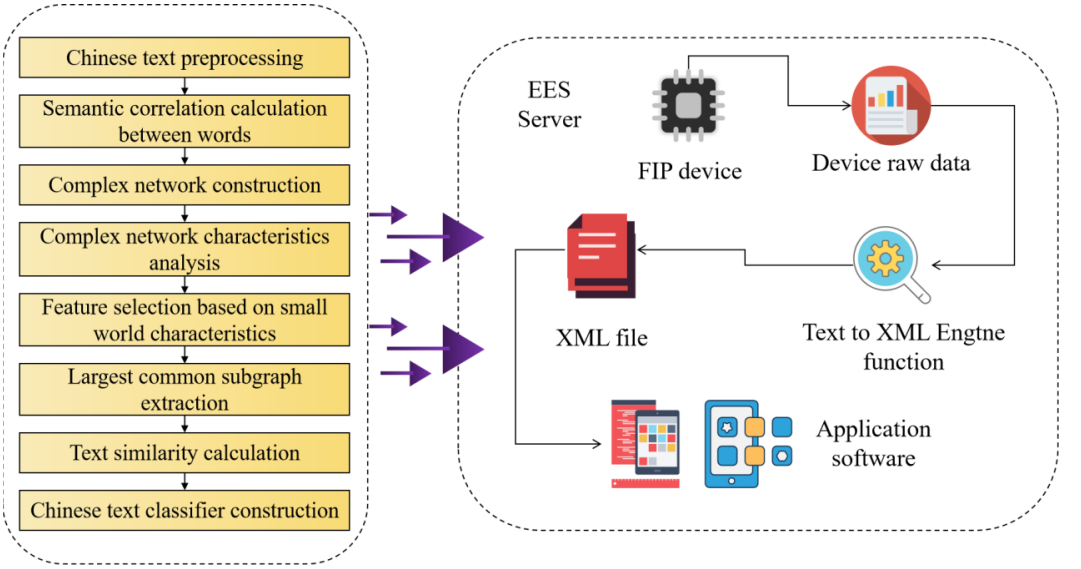
Feature Extraction Model of Dialogue Text

After the new text data is obtained, the information will be analyzed through the analysis system; the text information related to the event and unrelated to the event will be analyzed, and the related topics will be classified and analyzed. Finally, the analysis results will be reported according to the analysis. The process of model construction mainly includes data preparation; that is, the existing experience is transformed into a training sample set. Then, the identified event-related texts are marked, and the typical features are extracted to train the model. Next, the model is evaluated. The trained model can be used for model prediction. In the process of feature extraction to form feature vectors, the text to be analyzed is segmented first, and the word bag model and word vectors are combined to segment the text to be analyzed into word expressions. According to the comparison of feature words in the feature library, the corresponding feature vector expressions can be obtained. The query process of text content information is shown in Figure 2.

Vectorization of Information

After word segmentation, the text will eventually be transformed into feature vectors; that is, vectorization will be performed. The appropriateness of word selection directly determines the effectiveness of classification algorithms. At present, many feature extraction algorithms are based on machine learning. A common shortcoming of these algorithms is that they ignore that different terms can express the same or similar meanings; in other words, they ignore the differences between

Figure 2. Text content information query process



the vocabulary level and the concept level—that is, semantics (Lulham, et al., 2011). This defect causes the extracted entries to be inappropriate and reduces the precision of the classification effect index. The essence of a text vector is that the weights of feature words are weighted and added to obtain a new vector. Moreover, for texts belonging to the same category, their feature words are semantically related to each other, and their functions on classification are similar. These words are grouped together to represent the characteristics of this category.

Inverse document frequency indicates the distinctiveness of a vocabulary in the whole corpus, which is inversely proportional to the frequency of the vocabulary appearing in the whole corpus. The more times vocabulary appears, the less its ability to distinguish words, and the smaller the corresponding weight should be. The calculation formula is shown in formula (7):

$$idf_d = \lg\left(M \,/\, df_d\right) \tag{7}$$

Among them, M represents the total number of vocabularies in the corpus; df_d represents the number of vocabularies containing the word d in the corpus.

There are also many ways to perform matching between vectors. The most commonly used method is to use the cosine of the angle between the two vectors of the document and the query, that is, to calculate the cosine of the angle between two N -dimensional vectors in the vector space to measure the semantic similarity. The formulas are as follows:

$$Sim\left(D,Q\right) = \frac{D \cdot Q}{\left\|D\right\|\left\|Q\right\|} \tag{8}$$

$$D \cdot Q = \sum_{t=1}^{N} D_t Q_t \tag{9}$$

In Formulas (8) and (9), $D \cdot Q$ is the vector multiplication, and $\|D\|$ and $\|Q\|$ are the modulo of the vector. The forward variable $\alpha_t(i)$ is defined to represent the probability that the observation sequence at time t and the state of time t are s_i under the condition of given model λ, as shown in formula (10):

$$\alpha_t(i) = P(O_1, O_2, O_3, \ldots, O_t, q_t = s_i | \lambda) \tag{10}$$

The initial state can be obtained by sorting, as shown in formula (11):

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \tag{11}$$

Result calculation is shown in formula (12):

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{12}$$

Iterative recursion is shown in formula (13):

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad 1 \le t \le T-1, 1 \le j \le N \tag{13}$$

In spatial dimension reduction, feature selection and feature extraction can be used. Among them, feature selection can use chi-square statistics, cross entropy, document frequency threshold, and other methods to extract useful features in classification. Feature extraction can use cost analysis, linear discriminant analysis, and other methods to extract low-latitude features. By extracting features through spatial dimension reduction, the extracted features can be guaranteed to conform to the semantic description of the text.

If a text belongs to a certain category, the probability that its feature words are similar to the feature words in the feature set representing the characteristics of that category will be too high. Therefore, a set of features is considered to replace all the training texts in the same category, and the similarity between the test text vector and the group feature vector of each category is calculated. When selecting feature words in the text, the collection of entries that can represent the characteristics of the same category is completed and grouped together. After the group features are obtained, the similarity can be calculated only with each group feature. In this way, the possible categories are selected first, and then the similarity between the text to be tested and each training text in the categories is calculated by using the modified statistics, so that the category of the text to be tested is accurately obtained, and the whole classification is completed.

RESULT ANALYSIS AND DISCUSSION

Context framework is a three-dimensional space and a highly digital semantic structure, so it is a highly abstract text content. The writing unit of text is words, but there is no clear distinction mark between words. Therefore, the analysis of words is the first step in processing text information, and word segmentation is the basis of semantic analysis. The quality of word segmentation will have

a direct impact on the accuracy and recall rate of the final text classification. On the premise of semantic association, this paper first divides the words belonging to the same category in the text, and then merges those words with semantic association; that is, they are grouped together. When calculating the conditional entropy, the calculation of a single word is changed to the consideration of words belonging to the same group in a certain category. First, the sum of the total probabilities of this group of words in that category is calculated; then the conditional entropy of this group of words is calculated; and finally the information gain is obtained. This section verifies the proposed feature extraction model of dialogue text based on big data and machine learning by comparing it with manually labeled keywords and classical algorithms. Because the traditional text feature extraction algorithm has data loss, it needs a dialogue text feature extraction model for experimental analysis. The traditional text feature extraction algorithm has data loss because it does not consider the complex semantic features in the text. It can extract some features in the text only in a specific way, but cannot completely capture the complex semantic features in the text, resulting in data loss. The test results are shown in Figure 3.

It can be seen from Figure 3 that the proposed model basically converges when the training iteration reaches the 10th round, and the loss function value is only 0.25, so the training loss is small. This shows that the proposed model can achieve rapid convergence, and the training effect is ideal.

Test samples are respectively input into two evaluation models for testing, and the scatter diagram of predicted value and actual value of test samples tested by the ID3 algorithm model is shown in Figure 4. The scatter diagram of the predicted value and the actual value of the test sample tested by using the algorithm evaluation model in this paper is shown in Figure 5. The dots on the graph indicate the ratio of the predicted value to the actual value.

The improved weight formula not only pays attention to the distribution of terms in different categories, but also makes a semantic analysis of related terms in the same category and integrates them into the formula. The calculated weight provides richer information for text classification. Thus, the shortcomings of traditional word frequency anti-document frequency in feature extraction are resolved, the weight calculation formula is revised, and the verification experiment of text classification is finally completed. The model in this paper is compared with the ID3 algorithm model. The comparison results of recall rate are shown in Table 1 and Figure 6. The accuracy comparison results are shown in Table 2 and Figure 7. Text classification can be evaluated from the perspectives of classification result quality, complexity, algorithm simplicity, etc. The quality

Figure 3. Training loss

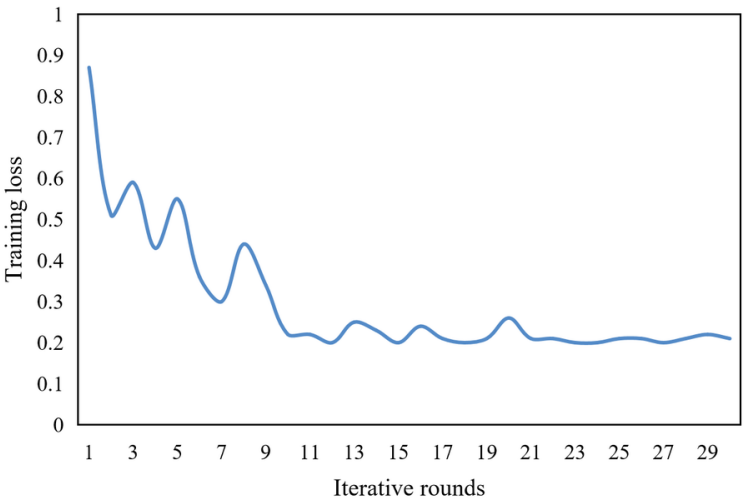


Figure 4. Scatter diagram of actual value and predicted value of ID3 algorithm

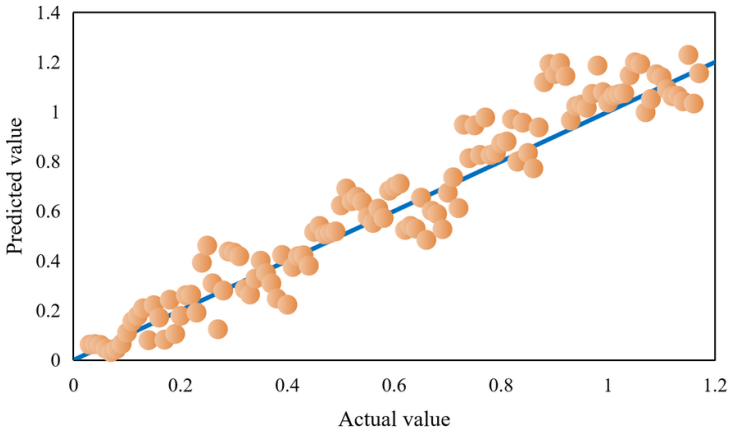
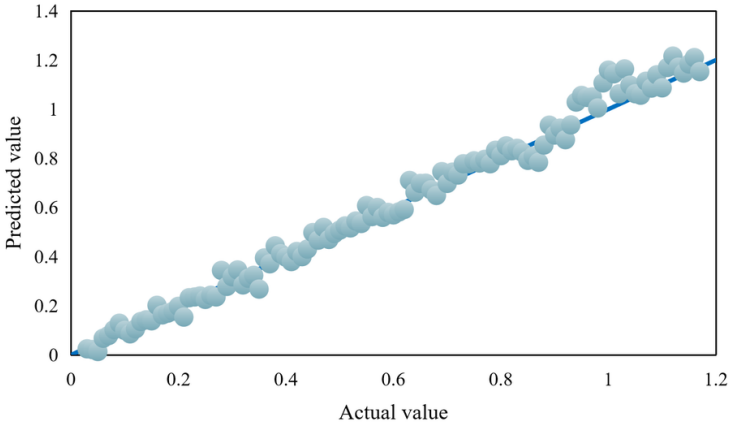


Figure 5. Scatter diagram of actual value and predicted value of this algorithm



of classification results is the most important. The term *effect* is usually used to collectively refer to the evaluation indicators for evaluating the quality of classification results. The evaluation indicators include Precision (P) and recall (R).

The results show that in most cases, the classification accuracy of this model is higher than 88% and the recall rate is higher than 85%; thus the goal of higher classification accuracy with less computation has been achieved. Based on the analysis of the experimental results, it is concluded that using the improved weight formula, both the accuracy rate and the recall rate have been improved to varying degrees, and the improvement is more evident in the two indicators of average accuracy rate and average recall rate. Experiments have shown that the semantic association between words within a category has a significant impact on the weight value, and introducing this factor into the formula can significantly improve the final classification effect. In this experiment, special processing was performed on the corpus; for instance, the proportion of synonyms in the document. The purpose of this approach is to make the categories more distinct and more conducive to data analysis. It has obvious advantages for specialized texts. There are a large number of semantically related words in Chinese information, so the method in this paper has great application value and will be of significant help in improving the effectiveness of text classification.

Table 1. Recall test results of the algorithm

Iterations	Recall	Iterations	Recall
0	0.901	55	0.941
5	0.945	60	0.946
10	0.864	65	0.852
15	0.874	70	0.87
20	0.91	75	0.953
25	0.853	80	0.943
30	0.904	85	0.931
35	0.902	90	0.927
40	0.899	95	0.929
45	0.861	100	0.866

Figure 6. Comparison of recall rates of different models

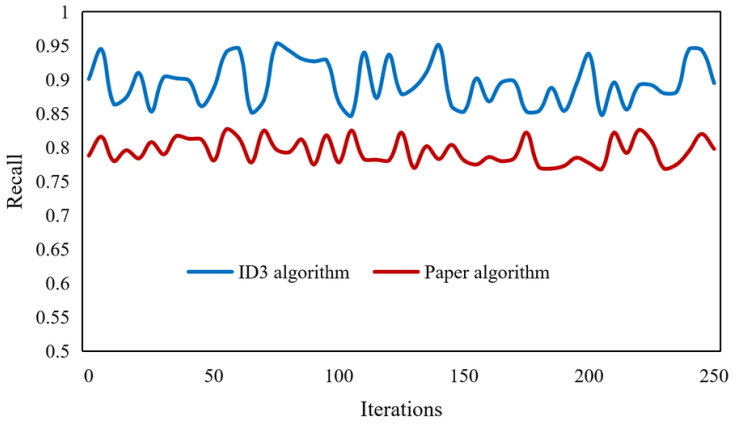
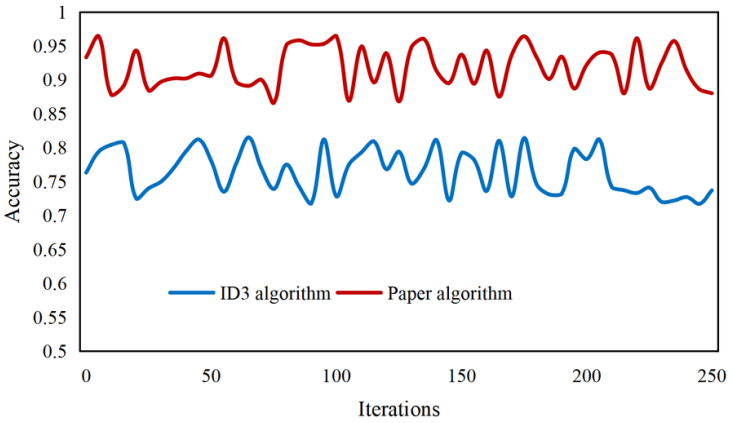


Table 2. Accuracy test results of the algorithm

Iterations	Accuracy	Iterations	Accuracy
0	0.763	55	0.735
5	0.794	60	0.777
10	0.804	65	0.815
15	0.807	70	0.77
20	0.725	75	0.739
25	0.74	80	0.775
30	0.75	85	0.743
35	0.77	90	0.718
40	0.795	95	0.812
45	0.812	100	0.728

Figure 7. Comparison of accuracy of different models



CONCLUSION

Text classification is the sorting of each text into a predefined category so that users can quickly and conveniently obtain the required information by some means of searching while reading the text. In this paper, a dialogue text feature extraction model based on big data and machine learning is constructed which transforms the high-dimensional space of text features into a low-dimensional space that is easy to process, so that the best feature words can be selected to represent the document set. The results show that in most cases, the classification accuracy of this model is higher than 88%, and the recall rate is higher than 85%; thus the model achieves the goal of higher classification accuracy with less computation. According to the calculation of the feature selection function, each feature word is given a different weight, and the best feature word is selected. Finally, the experimental results show that not only is the classification time greatly shortened, but the accuracy rate and recall rate are also significantly improved.

Although the research work in this paper has preliminarily confirmed the effectiveness of the proposed text feature extraction algorithm, there are still some aspects to be improved and perfected in practical application and experiments. The data set of this paper is small, and is stored mainly by text files. If the data volume is large, we must consider using a suitable database to manage the text, find out a suitable database, and improve the efficiency of text feature extraction.

DATA AVAILABILITY

The figures and tables used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This work was not supported by any funds.

ACKNOWLEDGEMENTS

The authors would like to show sincere thanks to those people who have contributed to this research.

REFERENCES

Barbantan, I., Porumb, M., Lemnaru, C., & Potolea, R. (2016). feature engineered relation extraction – Medical documents setting. *International Journal of Web Information Systems*, *12*(3), 336–358. doi:10.1108/IJWIS-03-2016-0015

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications*, *42*(6), 3105–3114. doi:10.1016/j.eswa.2014.11.038

Calvo, H., Paredes, J. L., & Figueroa-Nazuno, J. (2018). Measuring concept semantic relatedness through common spatial pattern feature extraction on EEG signals. *Cognitive Systems Research*, *50*(8), 36–51. doi:10.1016/j.cogsys.2018.03.004

Chakroborty, S., & Saha, G. (2010). Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification. *Speech Communication*, *52*(9), 693–709. doi:10.1016/j.specom.2010.04.002

Chambua, J., Niu, Z., Yousif, A., & Mbelwa, J. (2018). tensor factorization method based on review text semantic similarity for rating prediction. *Expert Systems with Applications*, *114*(12), 629–638. doi:10.1016/j.eswa.2018.07.059

Chapman, W. W., Savova, G. K., Zheng, J., Tharp, M., & Crowley, R. (2012). Anaphoric reference in clinical reports: Characteristics of an annotated corpus. *Journal of Biomedical Informatics*, *45*(3), 507–521. doi:10.1016/j.jbi.2012.01.010 PMID:22343015

Dong, J., Li, X., & Snoek, C. (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, *20*(12), 3377–3388. doi:10.1109/TMM.2018.2832602

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira, W. Jr. (2011). Word co-occurrence features for text classification. *Information Systems*, *36*(5), 843–858. doi:10.1016/j.is.2011.02.002

Garla, V. N., & Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, *45*(5), 992–998. doi:10.1016/j.jbi.2012.04.010 PMID:22580178

Jiang, Q., Wu, Z., & Kang, J. (2022). Semantic key generation based on natural language. *International Journal of Intelligent Systems*, *37*(7), 4041–4064. doi:10.1002/int.22711

Kang, H., & Youn, S. (2020). Performance analyses of different text feature extraction algorithms in restaurant fake review detection. *Transactions of the Korean Institute of Electrical Engineers*, *69*(6), 924–929. doi:10.5370/KIEE.2020.69.6.924

Karthikeyan, T., Sekaran, K., Ranjith, D., & Balajee, J. M. (2019). Personalized content extraction and text classification using effective web scraping techniques. *International Journal of Web Portals*, *11*(2), 41–52. doi:10.4018/IJWP.2019070103

Lee, L. H., Isa, D., Choo, W. O., & Chue, W. Y. (2012). High relevance keyword extraction facility for Bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications*, *39*(1), 1147–1155. doi:10.1016/j.eswa.2011.07.116

Lee, Y. K., Song, J., & Won, Y. (2019). Improving personal information detection using OCR feature recognition rate. *The Journal of Supercomputing*, *75*(4), 1941–1952. doi:10.1007/s11227-018-2444-0

Lin, Y. L., Brusilovsky, P., & He, D. (2011). Improving self-organizing information maps as navigational tools: A semantic approach. *Online Information Review*, *35*(3), 401–424. doi:10.1108/14684521111151441

Lulham, A., Bogacz, R., Vogt, S., & Brown, M. W. (2011). An infomax algorithm can perform both familiarity discrimination and feature extraction in a single network. *Neural Computation*, *23*(4), 909–926. doi:10.1162/NECO_a_00097 PMID:21222523

Lv, M., Chen, L., Chen, T., & Chen, G. (2018). Bi-view semi-supervised learning based semantic human activity recognition using accelerometers. *IEEE Transactions on Mobile Computing*, *17*(9), 1991–2001. doi:10.1109/TMC.2018.2793913

Ma, J., Chow, T., & Zhang, H. (2020). Semantic-gap-oriented feature selection and classifier construction in multilabel learning. *IEEE Transactions on Cybernetics*, *2020*(99), 1–15. PMID:32191902

Oskouie, P., Alipour, S., & Eftekhari-Moghadam, A. M. (2014). Multimodal feature extraction and fusion for semantic mining of soccer video: A survey. *Artificial Intelligence Review*, *42*(2), 173–210. doi:10.1007/s10462-012-9332-4

Qian, L., Zhou, G., & Zhu, Q. (2011). Employing constituent dependency information for tree kernel-based semantic relation extraction between named entities. *ACM Transactions on Asian Language Information Processing*, *10*(3), 73–96. doi:10.1145/2002980.2002985

Touzani, S., & Granderson, J. (2021). Open data and deep semantic segmentation for automated extraction of building footprints. *Remote Sensing (Basel)*, *13*(13), 2578. doi:10.3390/rs13132578

Vicient, C., Sanchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textual resources. *Engineering Applications of Artificial Intelligence*, *26*(3), 1092–1106. doi:10.1016/j.engappai.2012.08.002

Wang, F., Liu, Z., Zhu, H., Wu, P., & Li, C. (2021). An improved method for stable feature points selection in structure-from-motion considering image semantic and structural characteristics. *Sensors (Basel)*, *21*(7), 2416. doi:10.3390/s21072416 PMID:33915845

Wang, F., Xu, T., Tang, T., Zhou, M., & Wang, H. (2016). Bilevel feature extraction-based text mining for fault diagnosis of railway systems. *IEEE Transactions on Intelligent Transportation Systems*, *18*(1), 49–58. doi:10.1109/TITS.2016.2521866

Xie, J., Zhang, L., You, J., & Shiu, S. (2015). Effective texture classification by texton encoding induced statistical features. *Pattern Recognition*, *48*(2), 447–457. doi:10.1016/j.patcog.2014.08.014

Zhou, H., Han, A., Yang, H., & Zhang, J. (2019). Edge gradient feature and long-distance dependency for image semantic segmentation. *IET Computer Vision*, *13*(1), 53–60. doi:10.1049/iet-cvi.2018.5035