Research Article

The Application of Unsupervised Learning TF-IDF Algorithm in Word Segmentation of Ideological and Political Education

Silian Chen,1 Lina Wu,2 and Jiexian Zhuo3

1Ideological and Political Education, Guangxi Normal University for Nationalities, College of Marxism, Chongzuo, Guangxi 532200, China
2Guangxi Normal University for Nationalities, College of Marxism, Chongzuo, Guangxi 532200, China
3College of Physical Education, Guangxi Beibu Gulf University, Qinzhou, Guangxi 535011, China

Correspondence should be addressed to Silian Chen; 2016122555@jou.edu.cn

Received 30 May 2022; Revised 13 July 2022; Accepted 21 July 2022; Published 9 August 2022

Word segmentation is an essential part of Chinese learning. It often plays a decisive role in the quality of language processing results. Its domain vocabulary has the advantages of fast birth of new words and wide coverage. Large vocabulary, which has caused great difficulties in word segmentation and follow-up work in the field. This paper designs and implements a system. The system builds a corpus based on domain literature, trains a word-level language model based on statistical ideas, and uses the Viterbi algorithm to obtain preliminary Chinese word segmentation results. The word optimization algorithm optimizes the preliminary. The system provides users with functions such as keyword extraction, word frequency statistics, and word cloud drawing for the word segmentation results, so as to realize Chinese word segmentation and text analysis of field documents. The experimental results show that building system based on the corresponding ideas has improved the text analysis efficiency, related research on documents and text processing in the field have a certain promotion effect.

1. Introduction

Marxist theory and ideological and political education are usually called "two courses education" in the new educational system of universities. The establishment of this course provides a solution to students’ psychological contradictions. From the results of different two-course education, students need to use a large number of ideas to analyze cases and make in-depth summary in the learning process. Through the analysis of a large number of documents, this paper studies the related fields of the feasibility analysis scheme [1]. However, Chinese is significantly different from many Western Latin languages. In the written expressions of many Western languages, spaces have been used as fixed space symbols at the beginning of writing, Chinese is only words, sentences, and paragraphs, which can be simply delimited by obvious delimiters, but words do not have a formal delimiter, and words are relatively independent. However, in Chinese, the difference is that there is no fixed interval between the words, which also leads to the different segmentation of words in the same text sequence may directly lead to different meanings of the text sequence. Only by using the relevant knowledge of natural language processing can the text be further analyzed [2]. Therefore, which is often the basis and key natural language processing, and the result directly determines whether the subsequent results of natural language processing are ideal. At present, Chinese word segmentation can achieve high accuracy in the general field. However, in the face of word segmentation problems in many professional fields, general word segmentation methods often cannot achieve good word segmentation results. The characteristics of vocabulary and style have led to the following challenges in Chinese word segmentation based on supervised learning [3].

First, it requires interpretation of a large number of professional documents. Second, due to the relative lack of prior knowledge, the labeling effect of domain literature may not be ideal. The third is that due to the characteristics of the early starting point and fast development speed, the research direction is wide, the professional is very large, and the
general corpus is difficult to meet the needs of field document segmentation. Fourth, the update speed of research hotspots in the field of two-course education and the birth of new words are fast, and professional new words emerge in an endless stream, which brings certain challenges to corpora and manual annotation.

From the above research, we can know that the fundamental need of the two-course education is to pay attention to the research accuracy and efficiency analysis in this field. Combined with the relevant knowledge of "two-course" education in various states, this paper further analyzes the relevant information in this educational field. By supervising the learning contents of different subjects, the key information in the field of ideological education is constructed, to solve the need for manual labeling of labor costs, according to the characteristics of literature in related fields, through the Chinese word segmentation optimization algorithm to further improve the accuracy and through the keyword extraction algorithm which completes the extraction and provides functions such as to meet the needs of text analysis and to help the research work in the field.

2. State of the Art

Word segmentation in natural language has always been the focus of research in academia [4]. The research on Chinese word segmentation can be traced back to the 1980s. As early as more than 30 years ago, Professor Liang Nanyuan of Beihang University has considered the word segmentation as more than 30 years ago, Professor Liang Nanyuan of Beihang University has considered the word segmentation method of "checking the dictionary," that is, the sentence in the text to be segmented. Scan the dictionary from the beginning to the end. When the words in the dictionary are scanned, they are marked. When encountering a string that does not exist in the dictionary, it is divided into single-character words. This theory is enough to complete a simple Chinese word segmentation task [5]. In the process of this theoretical development, Dr. Wang Xiaolong from Harbin Institute of Technology further theorized this word segmentation method through research and summarized it as a word segmentation theory for the minimum number of words, that is, the optimal result of the sentence to be segmented. Sentences should be segmented into the smallest number of word strings [6]. In Chinese word segmentation-related technologies, in the 1990s, Professor Huang Changning from Tsinghua University put forward four different Chinese characters and search the dictionary corresponding to the number of Chinese characters each time. Although the code and explanation are different, the essence and result are the same. The longer the word, the higher the priority. This should be paid attention to in the research.

The main idea statistics is that the more two or more words appear adjacently in the corpus, the higher the probability that they should be segmented as a word [9]. According to the idea, the adjacent word combinations in the corpus and their corresponding occurrences are counted, and the word combination with the highest probability is used as the optimal word segmentation result under the corpus training model.

Li Liangjie combined the idea of statistics and the characteristics of the nature of the Semantic Web, proposed a Chinese word segmentation algorithm based on statistics and semantic information, and discussed the problem. Yuan Xianghui, on the other hand, used the conditional random field model unregistered words in addresses by analyzing the characteristics of Chinese words and expressions and independently built an address model knowledge base and designed an algorithm suitable for standardized word segmentation of Chinese addresses [10]. This also provides a great reference for Chinese word segmentation in specific fields.

3. Methodology

3.1. Text Mining and Classification. Text mining uses knowledge engineering technology, statistics, and machine learning algorithms to process text data such as documents and web pages stored in semistructured and unstructured forms and mines the potential connections between text content and semantic units and grammatical units, which can be used for text classification, cluster analysis, and other data mining and knowledge discovery processes [11]. Text mining is the mining knowledge from text content, and its general process includes data acquisition, text preprocessing, text representation, text mining analysis, and result evaluation, as shown in Figure 1 [12].

Text mining is an extension of data mining in large-scale text datasets. The acquisition of text data is the first step in text mining. Text corpus can be obtained by means of web crawling, text file reading, etc. [13]. The original text corpus is divided into isolated feature items by word segmentation. There are some meaningless and synonymous units in these feature items. The preprocessed text data needs to undergo feature selection, and these text feature items are represented as a unified data structure for subsequent analysis and modeling. Commonly used feature selections include filtering, wrapping, and embedding [14]. The purpose of feature extraction is to convert text data into numerical vectors in the vector space. Commonly used methods mainly construct the vectorized representation of text from the perspective of
word bag and word embedding. Then, for different text mining tasks, different algorithms need to be selected or designed to analyze and display the inherent information of the text, so as to extract the knowledge management process for specific tasks, such as text classification, cluster analysis, and trend prediction. Finally, it is necessary to select specific evaluation indicators according to the text mining task. For example, to evaluate the text classification model, indicators such as accuracy, precision, and recall can be selected, or the mining results can be displayed in the form of result visualization.

The study of text classification includes several subject areas, including natural language processing in linguistics, statistics in the field of classification mathematics in library science, and research topics such as pattern recognition, artificial intelligence, and neural networks in the computer field [15]. At first, text classification was carried out through expert rules (patterns), and an expert system was established by using knowledge engineering. The advantage of this is that it can solve the problem intuitively, but it is time-consuming and laborious, and the coverage and accuracy are limited [16].

Text classification is the task of mapping a binary \((d, c) \in D \times C\) to a Boolean value. The mapping is expressed mathematically as follows:

\[
\Phi : D \times C \longrightarrow \{T, F\}. \tag{1}
\]

Here \(D\) is the set of texts to be classified, \(C\) is the set of all predefined categories under a given classification system, \(D\) can be an infinite set, and \(C\) must be a finite set:

\[
d_i = (w_{i1}, w_{i2}, \ldots, w_{im}). \tag{2}
\]

Among them, \(d\) is the feature vector representing the text \(i\), and the cosine similarity is used to replace the distance metric in the text representation. The calculation formula of the similarity is as follows:

\[
\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \cdot \sum_{k=1}^{n} w_{jk}^2}}. \tag{3}
\]

Select the \(K\) training texts closest to the text \(d_i\) to be classified, that is, the \(K\) samples with the largest cosine similarity. The \(K\) here is an empirical value, ranging from tens to thousands. Generally, a suitable \(K\) value can be selected through cross-validation according to the distribution of the samples. By counting the text weights \(P(d, c)\) belonging to class \(c\) in the \(K\) training texts, we can further judge the predefined categories they belong to.

\[
P(d, c) = \sum_{d \in KNN(d)} \text{Sim}(d, d_j) \cdot f(d_j, c),
\]

\[
f(d_j, c) = \begin{cases} 1, & d_j \in c, \\ 0, & d_j \notin c, \end{cases} \quad \tag{4}
\]

\[
w_{it} = \frac{tf(t, d) \times \log (N/n_t + 0.01)}{\sqrt{\sum_{d \in D} [tf(t, d) \times \log (N/n_t + 0.01)]^2}}.
\]

3.2 Unsupervised TF-IDF Algorithm. Due to the different lengths of different documents, these frequencies have a large gap, so they need to be normalized so that these frequencies can be compared under the same environment [17].

There are many ways of feature weighting, which “equal weight” and “non-equal weight” [18]. “Equal weight” considers that feature items have the same importance in the entire training set, and they will not have any substantial impact on the classification results, so all feature items are given the same weight. In contrast, “nonequal weighting” believes that the importance of feature items is different, and the role of the main feature items can be enhanced. The current research has a high degree of acceptance is the “nonequilibrium weight” method, the most representative of which is the “IDF” weight [19–21].

3.3 Improved TF-IDF Algorithm

3.3.1 WF_TF-IDF Algorithm. In the TF-IDF algorithm, since different texts in the text library are usually manually labeled into several different categories, the algorithm only considers the total frequency of feature words in the short text library, ignoring the proportions in several categories.
Wang Gensheng et al. introduced the class-frequency variance into TF-IDF. The above method is recorded as WF_TF-IDF. The class frequency variance is calculated as follows:

$$\delta_i = \sqrt{\frac{1}{N} \sum_{j=1}^{n} \frac{(df(d_i, w_j)/Ndf(d_i, w_j))^2}{N}}.$$  \hspace{1cm} (5)

The size of the value indicates the degree of fluctuation of the word in the category. The larger the value, the more obvious the degree of fluctuation, the more unbalanced the distribution, and the stronger the discriminative effect on document classification. Therefore, the improved algorithm formula of TF-IDF is based on the following:

$$w(i, j) - \delta_{ij} = TF \times IDF \times \beta.$$  \hspace{1cm} (6)

Formula (6) is called WF_TF-IDF single-model feature representation, which can take into account the distribution of feature words in categories and text libraries at the same time.

In view of the shortcomings of the traditional TF-IDF algorithm, Zhu Meng combined CHI with the TF-IDF algorithm to form the CHI_TF-IDF single-model feature representation method. Generally, the distribution of characteristic words is considered from two perspectives: intraclass and interclass distribution. The former refers to the distribution probability of \(w_i\) between each text within the category, and the latter refers to the proportion of \(w\) in different categories of texts. Use \(\beta_{in}\) to represent the occurrence probability of \(w\) within a class and \(\beta_{out}\) to represent the interclass distribution of \(w\). The formulas are as follows:

$$\beta_{in} = \sqrt{\frac{\sum_{i=1}^{n} (P_{ih} - Q_h)^2}{n - 1}},$$  \hspace{1cm} (7)

$$\beta_{out} = \sqrt{\frac{\sum_{i=1}^{n} (P'_{ih} - Q'_h)^2}{m - 1}}.$$  \hspace{1cm} (8)

In formula (8), the coincidence represents the occurrences of each category \(z\) occurrences of feature words in all categories; it represents the mean of the proportion of the \(n\)th \(w\) in each category; \(m\) represents the total number of categories. They are expressed as follows:

$$P_{ih} = \frac{TF_{ih}}{TF_h},$$  \hspace{1cm} (9)

$$Q_h = \frac{P_{ih}}{n},$$  \hspace{1cm} (10)

$$P'_{ih} = \frac{TF_{izh}}{TF_h},$$  \hspace{1cm} (11)

$$Q'_h = \frac{P'_{ih}}{m}.$$  \hspace{1cm} (12)

In formula (9), \(TF_{ih}\) is the value of the \(h\)th feature word \(w\) in the category single text \(i\) and \(TF_{izh}\)th \(w\) in each document within the category. In formula (11), \(TF'_{izh}\) is the value of the total frequency of the \(h\)th \(w\) in all texts of category \(z\), which means the total of the \(h\)th \(w\) in all documents containing all categories. In general, the intraclass distribution and the interclass distribution show a different relationship with the expressiveness of the \(w\) classification, the former is inversely proportional to this relationship, and the latter is the opposite. In this section, we use the index \(\beta\) to represent the expressiveness of \(w\) being classified, and the formula is as follows:

$$\beta = \beta_0 \times (1 - \beta_{in}),$$  \hspace{1cm} (13)

$$w(i, j) = TF \times IDF \times \beta.$$  \hspace{1cm} (14)

It can be seen from formula (14) that the expressive power of \(w\) classification becomes stronger as the value of \(\beta\) increases and \(\beta\) in the classification of texts. The overall model block diagram is shown in Figure 2.

Although these commonly used methods can acquire features, they still exist. It is insufficient, resulting in poor classification effect. TF-IDF has been introduced in the previous section. The information gain is similar to its processing, which is to obtain the characteristics of the entire document. Without considering the occurrence of the feature words between categories, the feature vector containing the category difference cannot be well extracted. Mutual information and chi-square tests overestimate the role of low-frequency words in text classification, and believe that low-frequency words contain more information. The reality is that words that can express themes in a document may appear repeatedly.

In this section, Word2vec combined with the improved TF-IDF algorithm is recorded as WoTFI for feature acquisition. Word2vec can not only obtain low-dimensional word vectors, but also in the vector and the improved TF-IDF algorithm. It not only fully considers the importance of words in the text but also takes into account the differences in the within and between classes, so the combination. The theoretical support formula for feature selection is as follows:

$$w(i, j) - \delta_{ij} - \beta = TF \times IDF \times \delta_i \times \beta.$$  \hspace{1cm} (15)

4. Result

4.1. Analysis of the Demand for Word Segmentation

(1) The vocabulary formation method is special, and there are many hot words. Considering the needs of system application, the Chinese word segmentation algorithm should efficiently and accurately segment the domain vocabulary on the basis of meeting the basic needs. The word segmentation results should provide users with a function to further optimize the results
There are many research direction systems that may not meet the word segmentation needs of individual research directions, and users may have a large number of corpora related to research directions. The system should be able to standardize the corpus provided by users. Use the standardized corpus to train related models, and users should be able to independently add new words for different models, improve the model’s ability to recognize some words, and ensure the accuracy of the system’s word segmentation for individual research directions.

To extract the key information of the documents in the massive documents after word segmentation. Therefore, the system should have the function of document feature extraction, including keyword extraction, word frequency statistics, and other functions to summarize relevant literature information and provide users with research reference information.

Considering that the documents provided by users may not be plain text or plain text files, in order to deal with various types of documents provided by users and perform word segmentation, the system needs to identify documents of different file types and extract the text in the documents for word segmentation.

4.2. Implementation of Word Segmentation for Education Based on TF-IDF Algorithm. The hierarchical structure of “two-course” education is divided into four parts. The content includes four parts: presentation, business, support, and data. Figure 3 shows the overall structure.

1. The hierarchical panel of this module mainly shows the operating functions of various systems.

2. It carries out the business logic processing of the system, mainly including model customization, input and output management, Chinese word segmentation, and word segmentation.

3. This level of panel contains different background function support.

4.3. Word Segmentation Results. In the previous section, Chinese word segmentation and word segmentation optimization algorithms were implemented. Since there is no public test corpus in the research, this section is verified by a certain amount of labeled corpus in the test set constructed by self-labeling. Chinese word segmentation algorithm and word segmentation optimization algorithm are accurate. This section selects 200 annotated corpora and many more from the test set. The word segmentation results of each word segmentation scheme are compared to determine the pros and cons of different word segmentation schemes. The literature research directions of the test set are different, covering multiple research directions and hotspots.

N-gram is a commonly used language model in large vocabulary continuous speech recognition. For Chinese, we call it Chinese language model. The N-gram language types of language models according to the difference of N. Consider the characteristics of long vocabulary and complex structure in the field. The model is based on the assumption of
that the occurrence of the \( n \)th word is only related to the previous \( n-1 \) word and is not related to any other word. The probability of the whole sentence is the product of the occurrence probability of each word. If we know more about this sentence segment, we will get a more accurate answer. This tells us that the more the previous (historical) information, the stronger the constraint on the later unknown information. When using Google, enter one or several words, and the search box will usually provide several options like the following figure in the form of pull-down menu. These options are actually guessing the word string you want to search. Moreover, when you use the input method to input a Chinese character, the input method can usually contact a complete word. For example, if I input a “Liu,” the input method will usually prompt me whether I want to input “Liu Bei.” Through the above introduction, you should be able to keenly realize that this is actually based on the N-gram model. The specific comparison test results are shown in Figure 4. The three different colored columns represent different corpora.

In this paper, related experiments are also carried out on the influence of different amounts of training text on the classification model, and the results are shown in Figure 5. The four different feature selection algorithms increase with the increase in the number of training documents, because with the increase of training samples, the more information learned, the better the performance of the model training. Considering the presence and method of feature weight representation, the WF_TF-IDF and CHI_TF-IDF algorithms are better than the traditional TF-IDF, but the WoTFI model performs the best, and the WoTFI model has a very good classification effect. Within classes is considered in the weight calculation, and the text data information is used to the greatest extent. The quality of the text classification effect is mainly represented by the accuracy of text classification, so the improved WoTFI algorithm can represent the text well and obtain a high accuracy rate.

TF-IDF and its three weighted improved models, WF_TF-IDF, CHI_TF-IDF, and WoTFI, were tested. The support vector machine classification method tests the classification effects of the four models in different dimensions. The classification results of each model on the corpus are shown in Figure 6.

For the traditional TF-IDF and its three weighted models WF_TF-IDF, CHI_TF-IDF, and WoTFI, the classification effect increases with the increase of the dimension, and the accuracy continues to rise. When the dimension reaches 500, the classification effect reaches the best, then the dimension further increases, and the classification effect tends to be stable or slightly decreased. Therefore, appropriate dimension settings can achieve the best classification effect. If the dimension is too low, it may be difficult for the model to achieve the best result state, and the dimension is not as large as possible. If it is too large, many useless feature words are included, which affects the results and slows down the classification speed. Analyze the results of the improved classification model to classify the SST-1 data, as shown in Figure 7.

The analysis and fusion of the text representation model are affected by the single text representation model. It is
Figure 4: Comparison of Chinese word segmentation processing time in different order language models.

Figure 5: The effect of the number of texts on the word segmentation results.

Figure 6: Classification effects of four feature selection models in different dimensions.
necessary to set the most suitable dimensions for the two single models, WF_TF-IDF and CHI_TF-IDF, respectively. The results are shown in Figures 8 and 9.

As can be seen from Figure 8, when the dimension is 200, the effect of the WoTFI model changes with the change of WF_TF-IDF. When the dimension of the WF_TF-IDF model changes from 100 to 500, the classification result accuracy of WoTFI combined with SVM increases very fast. After the dimension is greater than 500, the increase is slow. When the dimension is too large, the classification effect decreases. As can be seen from Figure 9, when the WF_TF-IDF dimension is fixed to 200, the effect of the WoTFI model also changes with the change of the CHI_TF-IDF dimension. When the CHI_TF-IDF dimension increases from 100 to 500, the classification result accuracy growth rate is very fast. Later, as the dimension increases, the model training time becomes longer, the training speed becomes slower, and the effect is not obvious. Combining the two figures, the CHI_TF-IDF method is more stable than the WF_TF-IDF method.
5. Conclusion

(1) This section selects 200 annotated corpora and more corpora from the test set and compares the word segmentation results of each word segmentation scheme. According to the difference of N, N-gram language model takes into account the characteristics of long vocabulary and complex structure in this field. There are obvious differences in the processing time of Chinese word segmentation under different sequential language models.

(2) This paper also carries out relevant experiments on the impact of different numbers of training texts on the classification model. Four different feature selection algorithms increase with the increase of the number of training documents, because with the increase of training samples, the more information learned, the better the performance of model training. The WF_TF-IDF and CHI_TF-IDF algorithms are better than the traditional TF-IDF, and WoTFI model performs best. When the number of training documents exceeds 1400, the accuracy of WoTFI model exceeds 90%, WF_TF-IDF and CHI_TF-IDF. The accuracy of TF-IDF model is more than 85% and has remained stable since then. The accuracy of traditional TF-IDF model is lower than the first three. The improved WoTFI algorithm can represent text well and obtain high accuracy.

(3) TF-IDF and its three weighted improved models, WF_TF-IDF, CHI_TF-IDF, and WoTFI, were tested. The classification effect of the four models increases with the increase of dimension, and the accuracy continues to improve. When the dimension reaches 500, the classification effect reaches the best, and then, the dimension further increases, and the classification effect tends to be stable or slightly decreases. Therefore, proper dimension setting can achieve the best classification effect. When the dimension is 200, the effect of the WoTFI model changes with the change of WF_TF-IDF. When the dimension of the WF_TF-IDF model changes from 100 to 500, the classification result accuracy of WoTFI combined with SVM increases very fast. When the WF_TF-IDF dimension is fixed to 200, the effect of the WoTFI model also changes with the change of the CHI_TF-IDF dimension. When the CHI_TF-IDF dimension increases from 100 to 500, the accuracy growth rate is very fast. The classification results show that the algorithm model can not only capture the two-way semantic dependencies, process sequential text data of different lengths, and obtain deep text features by capturing the shape and morphology information of words or phrases but also effectively retain the semantic information of short text and avoid long text at the same time. There were gradient explosion and disappearance in sequence training. The word segmentation system based on the corresponding idea improves the efficiency of text analysis and promotes the related research.
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.

Acknowledgments

This study was funded by the Guangxi Normal University for Nationalities: A Comparative study of Ideal and Belief Education between Chinese and Vietnamese University students (2009JS002). The authors would like to show sincere thanks to those techniques who have contributed to this research.

References

[1] L. M. Zhang, “Comprehensive evaluation model of MOOC teaching quality of accounting major based on Rete algorithm,” in Proceedings of the International Conference on E-Learning, E-Education, and Online Training, pp. 100–110, 2021.

[2] E. A. Borja and M. Cerna, “Morphological and physicochemical characterization of seagrass species using unsupervised learning algorithm,” Journal of Science, Engineering and Technology (JSET), vol. 6, pp. 236–246, 2018.

[3] S. Peng, L. Wang, and Q. Xia, “The keyword extraction of Chinese medical web page based on WF-TF-IDF algorithm,” in Proceedings of the 2017 International conference on cyber-enabled distributed computing and knowledge discovery (CyberC), pp. 27–31, Nanjing, China, 2017.

[4] W. J. Zhang, X. X. Yang, Z. Z. Liu, and Y. L. Zhang, “Chinese word segmentation of ideological and political education based on unsupervised learning,” in Proceedings of the 2nd International Conference, pp. 110–113, 2019.

[5] S. Choudhury, L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, “Guess what moves: unsupervised video and image segmentation by anticipating motion,” pp. 91–93, 2022, arXiv e-prints.

[6] S. Gautam and J. Singhai, “Cosine-similarity watershed algorithm for water-body segmentation applying deep neural network classifier,” Environmental Earth Sciences, vol. 81, no. 9, pp. 27–31, 2022.

[7] G. J. Aguiar, R. G. Mantovani, S. M. Mastelini, A. C. de Carvalho, G. F. Campos, and S. B. Junior, “A meta-learning approach for selecting image segmentation algorithm - ScienceDirect [I],” Pattern Recognition Letters, vol. 128, pp. 480–487, 2019.

[8] H. Fan and Y. Qin, “Research on text classification based on improved TF-IDF algorithm,” in 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018), Atlantis Press, 2018.

[9] X. Hu and D. Li, “Research on a single-tree point cloud segmentation method based on UAV tilt photography and deep learning algorithm,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4111–4120, 2020.

[10] X. Huang and Q. Wu, “Micro-blog commercial word extraction based on improved TF-IDF algorithm,” in 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), Xi’an, China, 2013.

[11] D. Diermeier, J. F. Godbout, B. Yu, and S. Kaufmann, “Language and ideology in congress,” British Journal of Political Science, vol. 42, no. 1, pp. 31–55, 2012.

[12] R. Zheng and G. Zheng, “An artificial intelligence data mining technology based evaluation model of education on political and ideological strategy of students,” Journal of Intelligent and Fuzzy Systems, vol. 40, no. 5, pp. 1–12, 2020.

[13] G. Lin, “Study on innovative research on ideological and political education in institutions of high learning in the era of internet plus,” in Proceedings of the 2020 3rd International Seminar on Education Research and Social Science (ISERSS 2020), pp. 92–96, 2021.

[14] Y. Si, “On the influence of party history education on ideological and political education for medical students,” Journal of Higher Education Research, vol. 3, no. 2, pp. 141–145, 2022.

[15] Z. K. Hou, H. L. Cheng, S. W. Sun, J. Chen, D. Q. Qi, and Z. B. Liu, “Crack propagation and hydraulic fracturing in different lithologies,” Applied Geophysics, vol. 16, no. 2, pp. 243–251, 2019.

[16] Q. Wang, “Application of clustering algorithm in ideological and political education in colleges and universities,” Journal of Physics: Conference Series, vol. 1852, no. 3, p. 32041, 2021.

[17] D. Dessi, R. Helouai, V. Kumar, D. R. Recupero, and D. Riboni, “TF-IDF vs word embeddings for morbidity identification in clinical notes: an initial study,” pp. 92–100, 2021, https://arxiv.org/abs/2105.09632.

[18] Y. Wu, X. Yang, and Q. Song, A New Supervised Learning Algorithm of Recurrent Neural Networks and l2 Stability Analysis in Discrete-Time Domain, INTECH Open Access Publisher, 2018.

[19] J. Wei, H. Cheng, B. Fan, Z. Tan, L. Tao, and L. Ma, “Research and practice of ‘one opening-one closing’ productivity testing technology for deep water high permeability gas wells in South China Sea,” Fresenius environmental bulletin, vol. 29, no. 10, pp. 9438–9445, 2020.

[20] J. Han, H. Cheng, Y. Shi, L. Wang, Y. Song, and W. Zhang, “Connectivity analysis and application of fracture cave carbonate reservoir in Tazhong,” Science Technology and Engineering, vol. 16, no. 5, pp. 147–152, 2016.

[21] S. Liu and X. Ma, Eds., Advanced Hybrid Information Processing 5th EAI International Conference, ADHIP 2021, Virtual Event, October 22-24, 2021, Proceedings, Part II, Springer Nature, 2022.