Research on Intelligent Natural Language Texts Classification

Chen Xiao Yu[1]
Computer School
Beijing Information Science & Technology University
Beijing, China

Zhang Xiao Min[2]
Academy of Agricultural Planning and Engineering
Ministry of Agriculture and Rural Affairs
Beijing, China

***Abstract*—Natural language texts widely exist in many aspects of social life, and classification is of great significance to its efficient use and normalized preservation. Manual texts classification has the problems such as labor intensive, experience dependent and error prone, therefore, the research on intelligent classification of natural language texts has great social value. In recent years, machine learning technology has developed rapidly, and related researchers have carried out a lot of works on the texts classification based on machine learning, the research methods show the characteristic of diversification. This paper summarizes and compares the texts classification methods mainly from three aspects, including technical routes, text vectorization methods and classification information processing methods, in order to provide references for further research and explore the development direction of the texts classification.**

***Keywords*—Machine learning; natural language texts; text vectorization; classification information processing**

I. Introduction

Intelligent classification of natural language texts has many application scenarios in social life, and related research has important social value. In recent years, the research of natural language texts classification based on machine learning has received a lot of attention. Researchers tried to develop corresponding intelligent texts classification methods for different types of natural language texts, and obtained many valuable research results. The research objects involved the fields of government affairs, justice, energy and electricity, transportation, medical care and health, agriculture, science and technology, intellectual property, business, etc. The factors affecting the applicability of texts classification methods mainly include two points; firstly, the differences in the characteristics of different types of texts may affect the applicability of classification methods; secondly, the possible limitations of the training data used in model training might limit its applicability to other data of the same type.

From the perspective of technical route, the appropriate technical routes for different texts classifications are usually different due to the differences in text types and text samples. In general, the technical routes usually involve data collection, data preprocessing, text vectorization, classification information processing, classification and other links. In relevant research reports, through comparative analysis, the technical routes which were finally adopted and achieved good results were different in the links and the basic technical methods involved. For example, for Chinese texts, it is usually necessary to add word segmentation link before text vectorization; in some classification scenarios, the classification effect could be optimized by adding feature selection or feature extraction link to reduce dimensionality before text vectorization; in the classification information processing, there are also widely differences in the appropriate information processing levels for different types of text data.

From the perspective of specific technical method, the applicability of technical method to text types has the characteristics of difference and diversity. A method may have different applicability when targeting different types of texts. Zhu F. P. et al. has achieved good classification results in the classification of news texts in the shipbuilding industry using the method based on SVM [1]; Zhao M. et al. compared the classification effects of LSTM, SVM, and CNN methods in the classification of dietary health texts, the results showed that LSTM had the best effect in the corresponding scenario, which was better than SVM [2]. The same type of texts may be classified effectively based on different types of methods, though there might be some differences in the classification accuracy. For example, Bao X. et al. compared the classification effects of the multi-instance learning method, SVM and KNN in the classification of patent texts, and the results showed that the multi-instance learning method could achieve good classification results [3]; Wen C. D, et al. used the BIGRU method in the classification of patent texts, and also achieved good results [4].

Natural language texts classification methods usually include two core parts, one is the text vectorization method, and the other is the classification information processing method. Text vectorization refers to the representation of natural language text in the form of vectors for further data processing, the technical methods used in related research include word frequency-based methods, distributed static word vector-based methods, and distributed dynamic word vector-based methods and so on. The technical methods based on word frequency express text in the form of vector mainly by counting the frequency of word appearing in the text, the principle is relatively simple and the implementation is relatively convenient, but the context-related information could not be preserved in the text vectorization; the distributed word vector methods, which understand and represent words based on context information, could effectively retain context-related information in text vectorization; the distributed dynamic word vector methods are different from the static methods, the main

Funding Project: Beijing Information Science & Technology University-Computer School- Scientific Research Business Work Funds (5029923412).

difference is that the distributed dynamic word vector methods could distinguish the polysemy of a word in different contexts. Classification information processing is the process of further processing for the information in the vectors to obtain more accurate classification prediction information after texts are represented in the form of vectors. The technical methods involved in current related research include classical machine learning methods such as SVM, NB, the methods based on signal type of neural network, the methods based on vertical combination of multiple types of neural networks, the methods combining attention mechanism, the methods based on ensemble learning strategies, etc. In general, the methods used in current natural language texts classification research usually involve the combination of a variety of basic technologies, and with the continuous development of the basic technology and the breadth and depth of the texts classification research, the basic technologies and the combination strategies are becoming more and more diverse.

This paper analyzes and compares the research methods of natural language texts classification based on machine learning from three aspects: technical routes, text vectorization methods, and classification information processing methods. On the one hand, it is expected to provide reference support for subsequent related research and application, and on the other hand, it is expected to explore the technical development direction of this field based on the analysis of high-level research reports in recent years.

II. Technical Route of Intelligent Natural Language Texts Classification

The links involved in natural language texts classification include data acquisition, data preprocessing, splitting into words (for Chinese), text vectorization, feature selection, feature extraction and dimensionality reduction, classification information processing and classification. Among them, data preprocessing, splitting into words (for Chinese), text vectorization, classification information processing and classification are usually necessary links.

Natural language texts classification modeling involves multi-link collaboration. Data acquisition methods mainly include crawling data from internet, internal data and public data sets; data preprocessing mainly includes data desensitization, deduplication, removing invalid data, removing incomplete data, completing incomplete data, etc.; the link of splitting into words is usually based on jieba, and if it is combined with a thesaurus based on professional fields, better result might could be achieved; the methods commonly used in text vectorization mainly include BERT, word2vec, and TD-IDF; feature selection can reduce the vector dimension, the methods commonly used include chi-square test, removal of low-frequency words, etc.; the use of feature extraction and dimensionality reduction may improve the classification effect, related methods include PCA, etc.; classification information processing is used to find out the connection between the text content and the classification to which it belongs, related methods mainly include neural-based methods, ensemble learning-based methods in which, different types of technologies could be used in combination; in order to obtain classification results, the softmax technical method was commonly used.

The differences and diversity exist in the technical routes of the machine learning texts classification for different types of texts or different datasets. We summarize a technical route based on the commonality summary and difference supplement, which covers the main links involved in the texts classification, and in a specific study, the technical route could be used as a basis for further applicability deletion or adjustment, the technical route is shown in Fig. 1.

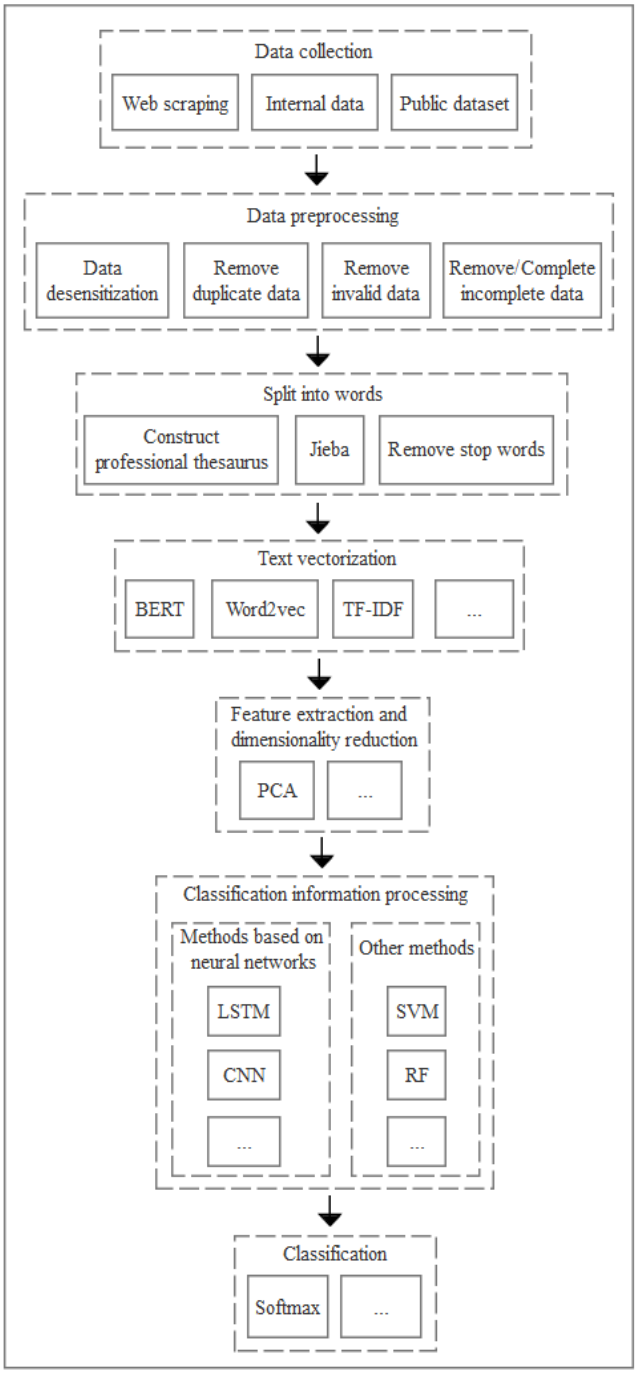

Fig. 1. The Technical Route of Texts Classification Research.

III. TEXT VECTORIZATION METHODS

Text vectorization is one of the core links of natural language texts classification. After preprocessing, text data usually needs to be expressed in vector form firstly, and then classification information processing would be operated. Text vectorization methods could be mainly divided into four categories: methods based on distributed dynamic word vectors, methods based on distributed static word vectors, methods based on topic and methods based on word frequency, as shown in Table I.

The technical methods based on word frequency include one-hot, TF-IDF, etc. One-hot is the simplest text vectorization technical method, whose basic principle is to use an N-dimensional vector to represent text based on the size of thesaurus, the N is the number of the words in the thesaurus, for a piece of text data, each dimension value of the N-dimensional vector corresponds to a word in the thesaurus, and the value range of the dimension variables is 0 or 1, if the word corresponding to a dimension variable appears in a piece of text data, its corresponding value would be 1, and if it does not appear, the corresponding value would be 0; one-hot is simple in principle and convenient in implementation, which also has many shortcomings, for example, when the scale of thesaurus is relatively large, the dimension of the vector will also be expanded accordingly, which is not conducive for data processing and classification. One-hot could be understood as the simplest word frequency-based text vectorization technical method, furthermore, TF-IDF is a kind of some more complex text vectorization technical method based on word frequency, compared to other technical methods, TF-IDF still has the characteristics of simple principle and convenient implementation, which is widely used in current text vectorization. Text vectorization methods based on word frequency could often achieve good application results in texts classification, which also have some important disadvantages, including: (1) the extraction of text features would ignore contextual information, (2) the position information where the word appears would be lost in text vectorization, etc.

The text vectorization methods based on distributed word vector have received extensive attention in recent years. The basic principle of distributed word vector method is to understand and represent word based on the context, therefore, compared with the methods based on word frequency, it could effectively solve the problem of the loss of contextual information in the vectorized representation of text. The context-based text vectorization methods could usually more effectively extract the classification information from text data, thereby improving the classification effect, and in different methods of this type, there are also differences in the scope of the context used, based on which the technical methods could be divided into partial text information-based methods and all text information-based methods; relatively, all text information-based methods usually have the advantages in relational information extraction.

Distributed word vector methods include distributed dynamic word vector methods and distributed static word vector methods. Distributed static word vector methods understand and represent words based on context, however, it cannot solve the problem of polysemy, that is, the same word may have different meanings in different contexts, in the text vectorization based on static word vector, a word could only have one representation, so it is impossible to distinguish the different interpretations of a word in different contexts. The main difference between the distributed dynamic word vector methods and the distributed static word vector methods is that the dynamic methods could distinguish the different interpretations of a word.

Topic-based text vectorization methods are another type of commonly used methods in natural language texts classification, including LDA (latent Dirichlet allocation), improved LDA, etc. The basic principle of LDA-based texts classification methods is to firstly extract topic information based on the texts, and then implement texts classification based on the topic information.

TABLE I. TEXT VECTORIZATION METHODS

No.	Category	Method	Using Examples	References
1	Distributed dynamic word vectors	BERT (bidirectional encoder representations from transformers)	Social e-commerce text, power grid equipment defect text	[5], [6]
		ALBERT	Patent text (ALBERT VS Word2vec, global vectors for word representation)	[4]
2	Distributed static word vectors	Word2vec	Railway signal equipment breakdown short text, healthy diet text (Word2vec VS term frequency-inverse document frequency, bag-of-words model), rice knowledge text (Word2vec VS one-hot, term frequency-inverse document frequency, hashing), ultra-short commodity text, news text	[7], [2], [8], [9], [10]
		CLW2V (character level Word2Vec)	Railway text (CLW2V VS term frequency-inverse document frequency, Word2vec)	[11]
		Fasttext	Coal mine accident case text	[12]
3	Topic	LDA (latent Dirichlet allocation)	Patent text	[13]
		MULCHI-labeled LDA (improved LDA)	Science and technology video text	[14]
4	Word frequency	Bag-of-words model	Hypertension medical record text	[15]
		TF-IDF (term frequency-inverse document frequency)	People's congress report text, cultural tourism text	[16], [17]

In summary, the methods used in text vectorization have developed from simpler methods such as ont-hot and word frequency-based methods, to the methods based on static word vectors, and then to the methods based on dynamic word vectors. Zhao M. et al. compared the text vectorization methods of word2vec, TD-IDF, and bag-of-words in their research on dietary health texts classification, and the research results showed that word2vec had the best effect in corresponding scenario [2]; Wen Ch. D. et al. compared ALBERT, word2vec, and glove text vectorization methods in their research on patent texts classification, and the results showed that the ALBERT method had the best effect [4].

Text vectorization is the basis of natural language texts classification, with the development of related technologies and the extensive and in-depth development of related research, the development of the text vectorization methods could be used in texts classification currently presents two trends, firstly, new technical methods emerge in an endless stream, the new technical methods could usually better retain classification-related information in text vectorization transformation, thereby providing a good foundation for accurate classification; secondly, in the classification of different types of texts, the applicable text vectorization methods reflect the diversified development trend, the emergence of new technical methods has enriched the options, and different technical methods have different characteristics and applicability, increasing the diversity of text vectorization technical methods.

IV. Classification Information Processing Method

Classification information processing is another core link of natural language texts classification. After the text data is represented in the form of vector, the including classification information needs to be extracted or processed to realize texts classification. The methods mainly include neural network methods, attention machine combined methods, ensemble learning methods, active learning methods and other methods, as shown in Table II.

TABLE II. Classification Information Processing Methods

No.	Category	Method	Using Examples	References
1	Neural network	LSTM (long-short term memory network)	Eating healthy text (LSTM VS SVM, CNN)	[2]
		BILSTM (bidirectional LSTM)	Electricity audit text	[18]
		CNN (convolutional neural network)	Electrical equipment defect text	[19]
		MCNN (multi-pooling convolutional neural network)	Short railway signal equipment failure text	[7]
		CCNN (combined CNN)	News text	[10]
		CNN-SVM (CNN-support vector machines)	Nursing adverse event text	[20]
		CRNN (improved convolutional recurrent neural network)	Police text	[21]
		MNN (multilayer neural network)	Government website mailbox text (MNN VS naive Bayes, random forest, decision tree)	[22]
		BIGRU (bidirectional gated recurrent unit)	Patent text	[4]
		Method based on DCNN (deep CNN)	Rice knowledge text (DCNN VS BILSTM, attention-BIGRU, RCNN, DPCNN)	[8]
2	Attention machine	BILSTM-attention	Power grid equipment failure text, railway traffic accident text, commodity text	[23], [24], [25]
		CNN-NLSTM-attention (method based on CNN, nested long-short term memory network and attention mechanism)	News text	[26]
		LS-GRU (an improved GRU deep learning framework)	Medical text (LS-GRU VS BIGRU, LSTM, GRU)	[27]
		ERNIE (enhanced representation from knowledge integration)	People's congress report text	[16]
3	Active learning	SVD-CNN combined with improved active learning (CNN based on singular value decomposition algorithm combined with improved active learning)	Barrage text	[28]
4	Ensemble learning	Beam search ensemble	Short medical text	[29]
5	Other methods	NB (naive Bayes)	Cultural tourism text, agricultural text	[17], [30]
		SVM (support vector machines)	Hypertension medical record text, ship industry news text	[15], [1]
		Deep random forest	Super short commodity text (deep random forest VS k-nearest-neighbor, decision tree)	[9]
		Method based on multi-instance learning framework	Patent text (multi-instance learning VS support vector machines, k-nearest-neighbor)	[3]

Neural networks are widely used in the classification information processing in natural language texts classification, including classical neural networks, improved neural networks for specific classification problems and the combined neural networks formed by stacking different types of neural networks. There are two main advantages of the neural network methods applied to the classification information processing. On one hand, the neural network methods could usually extract the deep internal correlation information between text content and category with high quality through black-box information processing, compared with the traditional methods such as NB and SVM, neural network methods could commonly analyze classification information deeply, and thus achieve better texts classification result. On the other hand, the neural network methods have the advantage of being more flexible, in dealing with practical classification problems, the applicability could take classic neural network models (such as CNN, RNN, etc.) as basis to make suitability adjustment, and better classification results could be achieved by constructing improved neural network models to better adapt to text characteristics; different types of neural networks could also be combined vertically, using the combined neural network for texts classification information processing could realize multi-level classification information processing and possibly achieve more effective classification information extraction; neural networks could be well combined with attention mechanism to improve the processing effect for classification information, this kind of methods have attracted wide attention in related research in recent years, which could often effectively improve the classification effect of natural language texts, the combination with attention mechanism further increases the flexibility and applicability of neural network methods. In general, the advantages of deep extraction of classification information and flexibility make the neural network methods the most usable basic methods in the classification information processing of natural language texts.

Except the neural network methods, classical machine learning methods such as RF and SVM still play an important role in natural language texts classification information processing, the advantages of traditional machine learning methods are mainly that they are more convenient to implement and the principles are relatively simple, at the same time, although different types of methods have general differences in information extraction capabilities, the diversity of natural language text characteristics makes traditional machine learning methods could also be more suitable for some specific classification scenarios, which could also possibly obtain better classification results.

Ensemble learning method is an important kind of methods suitable for the classification information processing of natural language texts, the basic principle of ensemble learning is to use different basic technologies to independently train multiple models, and combine the outputs of multiple independent models through a certain strategy to obtain the final classification output. The basic technical methods used in ensemble learning could be traditional machine learning methods such as NB, or relatively complex neural network methods such as BILSTM; ensemble learning could comprehensively apply the advantages and applicability of various types of machine learning technologies to improve the effect of texts classification. In the classification scene with diverse text features, based on ensemble learning strategy, to select different types of machine learning technologies reasonably as the basic methods could improve the applicability of the overall model and achieve high-quality texts classification result.

In summary, the methods used in the classification information processing have developed from basic methods such as NB, SVM, etc., to the methods based on neural networks, and then to attention combined methods, etc. Zhao M. et al. compared LSTM, SVM methods in the classification of diet and health texts, the results showed that LSTM has the best effect in corresponding scene [2]; Liu Y. et al. combined convolutional neural network (CNN), nested long short-term memory network (NLSTM) and attention mechanism in the study of news texts classification, and achieved good results using the CNN-NLSTM-Attention method [26]. In addition, the ensemble learning method could effectively combine the advantages of multiple classification models to improve the classification accuracy, which has received more attention in recent years [29]. Relevant theories and methods are becoming more and more abundant, further research and application could start from the texts characteristics to refer to research reports of the same type of texts for method selection or the basis for optimization, so as to efficiently and accurately implement texts classification modeling.

V. Text Types

Natural language texts classification commonly involves multiple links, in each link, the relevant basic technical methods have different applicability in different classification problems, and the main reason is the diversification of natural language text features; in texts classification research, the application could take the text types as a basis, and refer to existing related research reports to select technical methods or build a basis for the further improvement.

In order to provide convenient reference for subsequent relevant researchers, we have sorted up the research reports cited in this paper according to the text types of their research objects, which involves the fields of energy and environment, justice, government affairs, transportation, medical care and health, agriculture, science and technology, commerce and so on, as shown in Table III. This paper conduct analysis and research mainly from the perspectives of the technical routes of natural language texts classification and the main technical methods suitable for the core links, which does not comprehensively cover the text types already involved in natural language texts classification research, it is expected that subsequent researchers could make quick reference based on the text types table organized in this article or take it as a basis for further improvement.

TABLE III. TEXT TYPES OF NATURAL LANGUAGE TEXTS CLASSIFICATION

Field	Text type	References
energy and electricity	coal mine accident case text	[12]
	power information communication customer service system text	[31]
	power grid equipment text	[6], [23]
	electricity audit text	[18]
	power equipment defect text	[19]
transportation	railroad accident text	[24]
	railway text	[7]
	ship industry news text	[1]
medical and health	image report text	[27]
	medical short text	[29]
	nursing adverse event text	[20]
	hypertension medical record text	[15]
	eating healthy text	[2]
agriculture	rice knowledge text	[8]
	agricultural text	[30]
government affairs	people's congress report	[16]
	government website mailbox text	[23]
judicial	judgment document	[32]
	police text	[21]
Intellectual property	patent text	[4], [3], [13]
science & technology	academic paper	[33]
	science and technology video text	[14]
business	social e-commerce text	[5]
	super short commodity text	[9]
	commodity text	[25]
society	news	[10]
		[26]
		[34]
other fields	digital library text	[35]
	barrage text	[28]
	travel text	[17]

VI. CONCLUSION AND OUTLOOK

In general, the research on natural language texts classification received a lot of attention in recent years, and the research reports involved many important aspects of social life，including intellectual property, government affairs, justice, energy and electricity, transportation, medical care and health, agriculture, science and technology, commerce and so on, which constructed a favorable foundation for further research and applications. In terms of texts classification methods, the methods show diversification, and different methods commonly have different characteristics and applicability differences, which are closely related to text characteristics; follow-up research could take text characteristics as the basis to select and further improve classification methods.

The core links in natural language texts classification include text vectorization and classification information processing, in recent years, the theories and technologies in these both two aspects have made great progress; the text vectorization methods have developed from the basic methods such as the methods based on word frequency to the methods based on distributed static word vectors, and then to the methods based on distributed dynamic word vectors; the classification information processing methods have developed from the methods based on a signal technology such as SVM, CNN or RNN to the methods based on vertical integration of multiple technologies and the methods based on ensemble learning. The emergence of new theories and technologies provide more options for subsequent texts classification research, and also constructed more advantageous basis for

further development of related theories and technologies; and at the same time, due to the diversification of text characteristics, the development trend of texts classification methods could still be diverse development.

In terms of similar studies comparison, the text classification modeling based on machine learning has received extensive attention in recent years, and at the same time, researchers have carried out some review and summary works. The relevant researches mainly summarized and analyzed from the aspects of research progress and development trends, and there are few systematic analyses for the links involved in text classification. This paper systematically analyzed the links involved in natural language text classification research, and summarized main research methods based on the core links. It's expected that this paper could efficiently provide reference for subsequent related research.

REFERENCES

[1] Zhu Fang Peng, Wang Xiao Feng, Text classification for ship industry news [J], Journal of Electronic Measurement and Instrumentation, 2020, 34 (01): 149-155.

[2] Zhao Ming, Du Hui Fang, Dong Cui Cui, Chen Chang Song, Diet health text classification based on word2vec and LSTM [J], Transactions of the Chinese Society for Agricultural Machinery, 2017, 48 (10): 202-208.

[3] Bao Xiang, Liu Gui Feng, Yang Guo Li, Patent text classification method based on multi-instance Learning [J], Information Studies: Theory & Application, 2018, 41 (11): 144-148.

[4] Wen Chao Dong, Zeng Cheng, Ren Jun Wei, Zhang Yan, Patent text classification based on ALBERT and bidirectional gated recurrent unit [J], Journal of Computer Applications, 2021, 41 (02): 407-412.

[5] Li Ke Yue, Chen Yi, Niu Shao Zhang, Social E-commerce text classification algorithm based on BERT [J], Computer Science, 2021, 48 (02): 87-92.

[6] Tian Yuan, Yuan Ye, Liu Hai Bin, Man Zhi Bo, Mao Cun Li, BERT pre-trained language model for defective text classification of power grid equipment [J]. Journal of Nanjing University of Science and Technology, 2020, 44 (04): 446-453.

[7] Zhou Qing Hua, Li Xiao Li, Research on short text classification method of railway signal equipment fault based on MCNN [J]. Journal of Railway Science and Engineering, 2019, 16 (11): 2859-2865.

[8] Feng Shuai, Xu Tong Yu, Zhou Yun Cheng, Zhao Dong Xue, Jin Ning, et al. Rice knowledge text classification based on deep convolution neural network [J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52 (03): 257-264.

[9] Niu Zhen Dong, Shi Peng Fei, Zhu Yi Fan, Zhang Si Fan, Research on classification of commodity ultra-short text based on deep random forest [J]. Transactions of Beijing Institute of Technology, 2021, 41 (12): 1277-1285.

[10] Zhang Yu, Liu Kai Feng, Zhang Quan Xin, Wang Yan Ge, Gao Kai Long, A combined-convolutional neural network for Chinese news text classification [J]. Acta Electronica Sinica, 2021, 49 (06): 1059-1067.

[11] Lu Bo Ren, Hu Shi Zhe, Lou Zheng Zheng, Ye Yang Dong, Character-level feature extraction method for railway text classification [J]. Computer Science, 2021, 48 (03): 220-226.

[12] Yan Yan, Yang Meng, Zhou Fa Guo, Ge Yi Fan, Comparison of text classification methods of coal mine accident cases based on Fasttext network [J]. Coal Engineering, 2021, 53 (11): 186-192.

[13] Liao Lie Fa, Le Fu Gang, Zhu Ya Lan, The Application of LDA Model in Patent Text Classification [J]. Journal of Modern Information, 2017, 37 (03): 35-39.

[14] Ma Jian Hong, Fan Yue Xiang, Science and technology video text classification based on improved labeled LDA model [J]. Computer Engineering, 2018, 44 (09): 274-279.

[15] Hu Jing, Liu Wei, Ma Kai, Text categorization of hypertension medical records based on machine learning [J]. Science Technology and Engineering, 2019, 19 (33): 296-301.

[16] Yu Hang, Li Hong Lian, Lü Xue Qiang, Text classification of NPC report contents [J], Computer Engineering and Design, 2021, 42 (06): 1772-1778.

[17] Wang Xiang Xiang, Fang Hui, Chen Chong Cheng, Classification technique of cultural tourism text based on naive Bayes [J]. Journal of Fuzhou University (Natural Science Edition), 2018, 46 (05): 644-649.

[18] Chen Ping, Kuang Yao, Hu Jing Yi, Wang Xiang yang, Cai Jing. Text categorization method with enhanced domain features in power audit field [J]. Journal of Computer Applications, 2020, 40 (S1): 109-112.

[19] Liu Zi Quan, Wang Hui Fang, Cao Jing, Qiu Jian, A classification model of power equipment defect texts based on convolutional neural network [J]. Power System Technology, 2018, 42 (02): 644-651.

[20] Ge Xiao Wei, Li Kai Xia, Chen Ming, Text classification of nursing adverse events based on CNN-SVM [J]. Computer Engineering & Science, 2020, 42 (01): 161-166.

[21] Wang Meng Xuan, Zhang Sheng, Wang Yue, Lei Ting, Du Wen, Research and application of improved CRNN model in classification of alarm texts [J]. Journal of Applied Sciences, 2020, 38 (03): 388-400.

[22] Wang Si Di, Hu Guang Wei, Yang Si Yu, Shi Yun, Automatic transferring government website e-mails based on text classification [J]. Data Analysis and Knowledge Discovery, 2020, 4 (06): 51-59.

[23] Tian Yuan, Ma Wen, Attention-BiLSTM-based fault text classification for power grid equipment [J]. Journal of Computer Applications, 2020, 40 (S2): 24-29.

[24] Han Guang, Bu Tong, Wang Ming Ming, Zheng Hai Qing, Sun Xiao Yun, et al. Text classification of railway traffic accidents based on dual-channel bidirectional long short term memory network [J]. Journal of the China Railway Society, 2021, 43 (09): 71-79.

[25] He Bo, Ma Jing, Li chi, Research on commodity text classification based on fusion features [J]. Information Studies: Theory & Application, 2020, 43 (11): 162-168.

[26] Liu Yue, Zhai Dong Hai, Ren Qing Ning. News text classification based on CNLSTM model with attention mechanism [J]. Computer Engineering, 2019, 45 (07): 303-308+314.

[27] Li Qiang, Li Yao Kun, Xia Shu Yue, Kang Yan. An improved medical text classification model: LS-GRU [J]. Journal of Northeastern University (Natural Science), 2020, 41 (07): 938-942+961.

[28] Qiu Ning Jia, Cong Lin, Zhou Si Cheng, Wang Peng, Li Yan Fang. SVD-CNN barrage text classification algorithm combined with improved active learning [J]. Journal of Computer Applications, 2019, 39 (03): 644-650.

[29] Zhang Bo, Sun Yi, Li Meng Ying, Zheng Fu Qi, Zhang Yi Jia, et al. Medical text classification based on transfer learning and deep learning [J]. Journal of Shanxi University (Natural Science Edition), 2020, 43 (04): 947-954.

[30] Zhao Yan, Li Xiao Hui, Zhou Yun Cheng, Zhang Yue. A study on agricultural text classification method based on naive bayesian [J]. Water Saving Irrigation,2018(02):98-102.

[31] Yu Xue Hao, Zhao Zi Yan, Ma Ying Long, Zheng Rong Rong, Xi Zi Yue, et al. Multi-label text classification for power ICT custom service system based on binary relevance and gradient boosting decision tree [J]. Automation of Electric Power Systems, 2021, 45 (11): 144-151.

[32] Weng Yang, Gu Song Yuan, Li Jing, Wang Feng, Li Jun Liang, Li Xin, Paragraph context-based text classification approach for large-scale judgment text structuring [J]. Journal of Tianjin University (Science and Technology), 2021, 54 (04): 418-425.

[33] Xue Feng, Hu Yue, Xia Shuai, Xu Jian Dong, Research on short text classification based on paper title and abstract [J]. Journal of Hefei University of Technology (Natural Science), 2018, 41 (10): 1343-1349.

[34] Hu Yu Lan, Zhao Qing Shan, Chen Li, Niu Yong Jie, A Fusion network for Chinese news text classification [J]. Journal of Chinese Information Processing, 2021, 35 (03): 107-114.

[35] Xu Tong Yang, Yin Kai. Text classification of digital library based on deep learning [J]. Information Science, 2019, 37 (10): 13-19.