Article | Received 12 October 2023; Accepted 28 November 2023; Published 10 January 2024
https://doi.org/10.55092/pcs20240002

Text classification study of enterprise support policies based on Bert+DPCNN+BiGRU model

Haibo Wang[1], Ningpeng Jiang[2,*], Ruzhi Xu[3], Shiyu Ma[1] and Yang Li[1]

[1] University of Jinan, Jinan 250022, China
[2] Shandong Science and Technology Service Development Promotion Center, Jinan 250101, China
[3] Shandong Fintech Institute, Jinan 250306, China

* Correspondence author; E-mail: whb2021875@163.com.

Abstract: To address the current problems of lack of dataset, unclear classification and unbalanced samples in policy text classification, we propose a text classification model of enterprise support policy based on BERT, deep pyramidal convolutional neural network (DPCNN) and bi-directional gated recurrent unit (BiGRU) - BGC. first, we use First, BERT is used to encode the dataset; then the improved DPCNN and BiGRU are used to extract features, and finally the features of the extracted sentences are combined and Softmaxed to obtain the final result. In order to make the BGC model more suitable for application in the field of enterprise support policy text classification, this paper improves the DPCNN from the enterprise perspective by adding its convolution layer to obtain deeper semantic information; in addition, this paper also improves the loss function and introduces FocalLoss to solve the problem of unbalanced samples and difficult samples in the dataset. The experimental results show that the accuracy rate is improved by 2 percentage points compared with the traditional BERT model, which effectively alleviates the problem of information overload of massive corporate support policies.

Keywords: business support policy; text classification; BERT model; difficult sample processing

1. Introduction

With the rapid development of China's economy, small and microenterprises have begun to emerge and are gradually becoming the mainstay of society as a whole. In order to promote the development of MSMEs, China has also introduced a series of supportive policies. For micro and small enterprises, it's a great time to enjoy China's policy dividend, however, the frequent policy releases make the policy text data grow dramatically and increase the labor cost [1]. After analysis, this paper finds several pain points for enterprises to pay attention to

Copyright©2024 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

policies: first, policies are distributed on different websites with complicated contents, and enterprises can not obtain relevant information in time; second, policy information released by the government is not completely for a certain type of enterprises, and may be mixed with other information, and enterprises will spend a certain amount of time cost when reading; finally, enterprises lack reasonable technical means to find the policy information that meets their own Finally, enterprises lack reasonable technical means to find policy information that meets their needs. Therefore, from the point of view of enterprises, especially small and micro enterprises, there is an urgent need for an automatic classification model to improve the efficiency of searching and matching, so that enterprises can enjoy the dividends in a timely manner.

The Guiding Opinions on Strengthening the Construction of Digital Government issued by China in 2022 pointed out that "building a scientific and standardized system of rules for the construction of digital government, and leading the overall situation of digital development with the construction of digital government" [2], which can be seen in China's determination and confidence in the construction of digital government. The rapid development of natural language processing, deep learning, image recognition, recommender systems and other technologies has also successfully helped this process, and enterprises wanting to get the information they want from the digital government must master effective automatic classification means. However, in the field of policy text mining, the application of related technologies is still in the exploratory stage, and most researchers study the policy text as a whole and do not specialize in the content part of the text.

In summary, this paper takes enterprises as the starting point, and the main work is: 1) to build a self-built enterprise support policy dataset, collect and clean the enterprise policy data released by China Government. 2) to improve the DPCNN module and loss function through data analysis, and finally propose the BGC model. 3) to compare with the traditional BERT [0] model, and confirm that the model can not only be more rapid accurate text classification for business support policies, but also greatly reduces labour costs.

2. Related studies

Text classification is the process of automatically classifying and labeling text data according to the author's intent using computers by using NLP, machine learning and other techniques [0]. Traditional methods are some statistical based models such as plain Bayesian, K-nearest neighbor and support vector machine models with simple algorithms but require a lot of feature engineering and relatively poor classification results [0]. Compared to traditional models, deep learning models not only reduce the labor cost of feature engineering, but also have relatively complex model structures and greatly improved learning capabilities. With the development of deep models, various complex models are also widely used in NLP: for example, neural network language models such as CNN, RNN, and LSTM have been effectively improved in text classification compared with traditional models [0], especially various preprocessing models, such as ELMo, BERT, and other general pre-trained language models, which appeared in 2018, have made experiments on NLP tasks like text classification

effects have been significantly improved [0]. In the field of policy text classification, Qian Li *et al.* [8] constructed a new energy industry policy classification model and compared it with two other commonly used text classification models, proving that the BERT model is excellent in terms of classification effect. loni Hagen *et al.* [0] analyzed the Federal Reserve policy statement text was analyzed and LDAMallet topic model approach was selected for topic mining, which significantly improved the classification and interpretation of the Fed's monetary policy statements. ZQ Shen *et al.* [0] analyzed for science and technology related support policy texts, first fused the extracted policy text keywords, and then input the fused results into the BERT model for optimization experiments, which led to a significant improvement in the classification accuracy of science and technology support policy texts.

After analyzing various support policies, we found that semantic understanding of policy texts [11] is very important. Since the BERT model has strong semantic understanding of text, better access to the semantic representation of sentences during training [12], and performs well in previous NLP tasks, we chose this model for our experiments. Zhao Dandan *et al.* [13] applied the BERT model to text classification and proved the superiority of the BERT model through comparison experiments with traditional algorithms such as traditional CNN, RNN, LSTM, *etc.* Yi Liu [14] used the bi-directional feature of the BiGRU model to obtain a more accurate analysis of e-commerce product quality reviews. Therefore, this study starts from the text of support policy, breaks the limitation of previous researchers who regarded the article as a whole, divides the content of the body of the policy into text by paragraphs, and then combines the BGC deep learning fusion model with the 10w text content of the support policy to carry out the classification study, extracts 4 target categories for experiments, and realizes the automatic classification of the text of the support policy.

3. Related theory and technology

3.1. BERT word embedding

BERT is a language representation model proposed by Google in 2018, which has achieved extraordinary results in 11 NLP tasks such as named entity recognition, SQuAD Q&A, and text classification once it was proposed. The BERT model not only learns the representation of words in context, but also learns sentence-level representation by next sentence prediction. It is also because of these two innovations that the BERT model has become a commonly used model in NLP. In this paper, we mainly use the result of policy segmentation as the input of BERT model, and then we finally get rich word embedding information with the help of prior information obtained by BERT using a large number of corpus training.

3.2. BiGRU model

BiGRU is developed from RNN, which is mainly used for sequential data processing tasks. In order to solve the problem of gradient disappearance and gradient explosion of RNN in long distance, LSTM was born, however, LSTM is not perfect, its training speed is very slow, which also brings great limitation to the training. The model effect is comparable to that of

LSTM. However, the one-way GRU cannot take into account the contextual information, so the model used in this paper is BiGRU.

4. BGC model

4.1. Improvements to DPCNN

First of all, after segmenting the text for the dataset, it is easy to observe that the text fields are relatively long. For this feature, the DPCNN is improved by adding a convolutional layer, which enables the model to learn deeper information.

The improved DPCNN is shown in Figure 1. Firstly, after the data is input, the data is region embeding, which transforms the data into sentence vectors for representation, and after representation, these feature vectors are input to the convolutional groups in turn, and each convolutional group contains a convolutional layer and the residuals connected from input to output. And each group is connected by a pooling layer with a step size of 2. Multiple convolutional groups can be set up to extract features until the final convolutional group gets the feature vector representation of the final text.

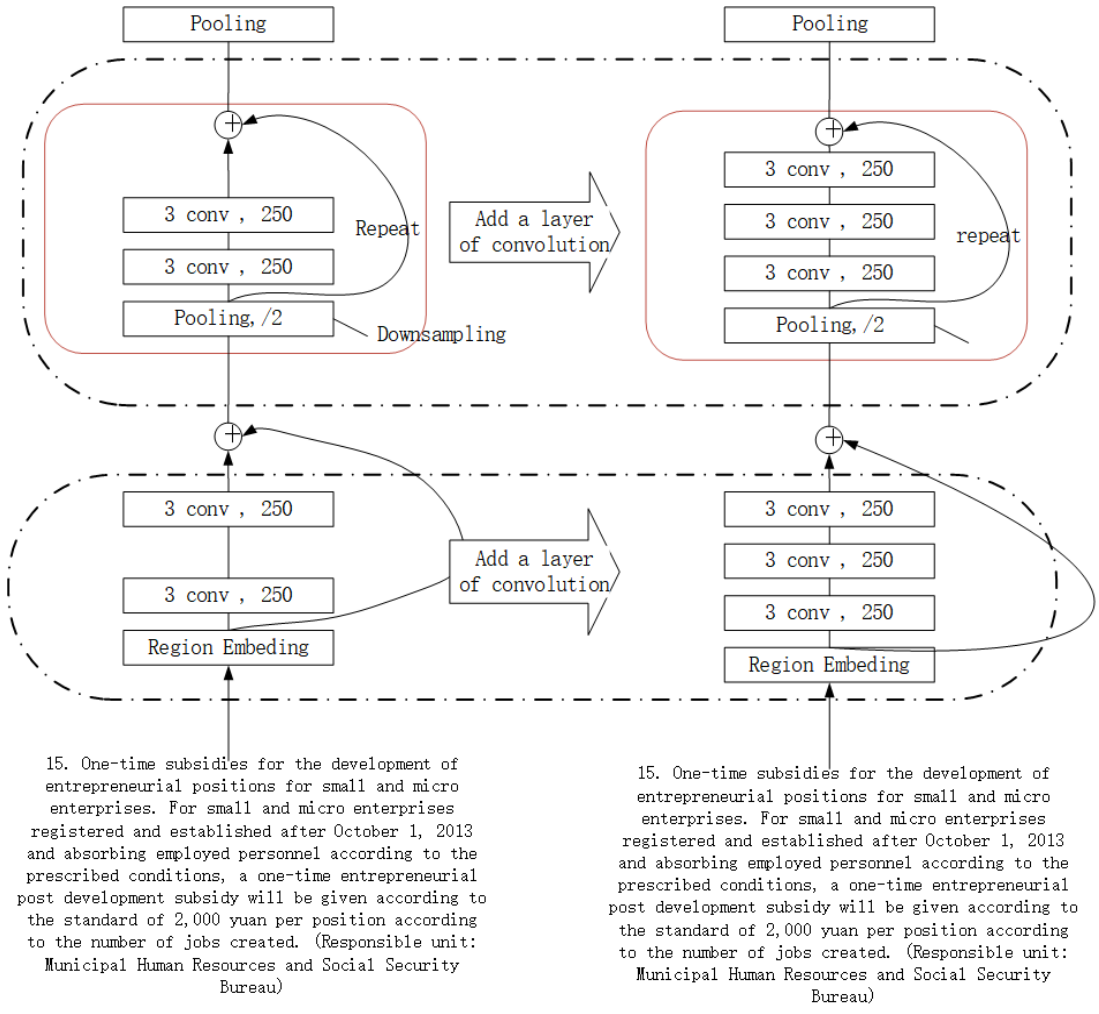


Figure 1. Comparison between DPCNN and DPCNN upgrade.

4.2. Improved loss function

The effect of traditional cross-entropy loss experiments is very unsatisfactory, and Zijiang Zhu [15] used the Focal Loss function to solve the problem of sample sentiment category imbalance,the effect is greatly improved by using this Loss in this paper. Table 1 shows the tuning process of FocaLoss. In this paper, it is observed from the labeled data that increasing the propor tion of tail labels in the training loss. increasing the proportion of incentive policies, subsidy policies, and individual-related policies in the training loss, is very effective in improving the overall classification accuracy of the model.The FL balance loss

Function is defined as shown in equation (1):

$$FL(\alpha,\gamma) = -\alpha(1-Pt)^{\gamma} * \log(Pt) * Samplesize \quad (1)$$

The function essentially introduces two variables α and β to dynamically adjust the original cross-entropy loss function to solve the problem of sample imbalance and difficult samples, and finally achieve an effective improvement of the classification effect. According to several experiments, the model effect is better when α=[0.4, 0.2, 0.2, 0.2] and γ = 2. The Focalloss tuning process is shown in Table 1.

Table 1. FocalLoss tuning process.

Model	α	γ	Accuracy
Fixed γ modifies α	[0.05,0.35,0.35,0.25]	2	0.9402
	[0.25,0.25,0.25,0.25]	2	0.9804
	[0.4, 0.2, 0.2, 0.2]	2	0.9929
Fixed α modifies γ	[0.4, 0.2, 0.2, 0.2]	1	0.9833
	[0.4, 0.2, 0.2, 0.2]	2	0.9929
	[0.4, 0.2, 0.2, 0.2]	3	0.9854

4.3. BGC model

The research framework of enterprise support policy text classification based on BERT deep learning model is shown in Figure 2 below, the whole framework can be divided into three parts: enterprise support policy text data collection and pre-processing, enterprise support policy text segmentation, and enterprise support policy text classification training. Firstly, we collect the enterprise support policy text data, determine the four classification dimensions, and then clean the enterprise support policy text data; next, we form the sample data set after text segmentation and annotation of the support policy text content; finally, we divide the whole data set into training set, validation set and test set in the form of 8:1:1, build the BGC model for experiments, and compare the experimental results. The results are compared and analyzed.

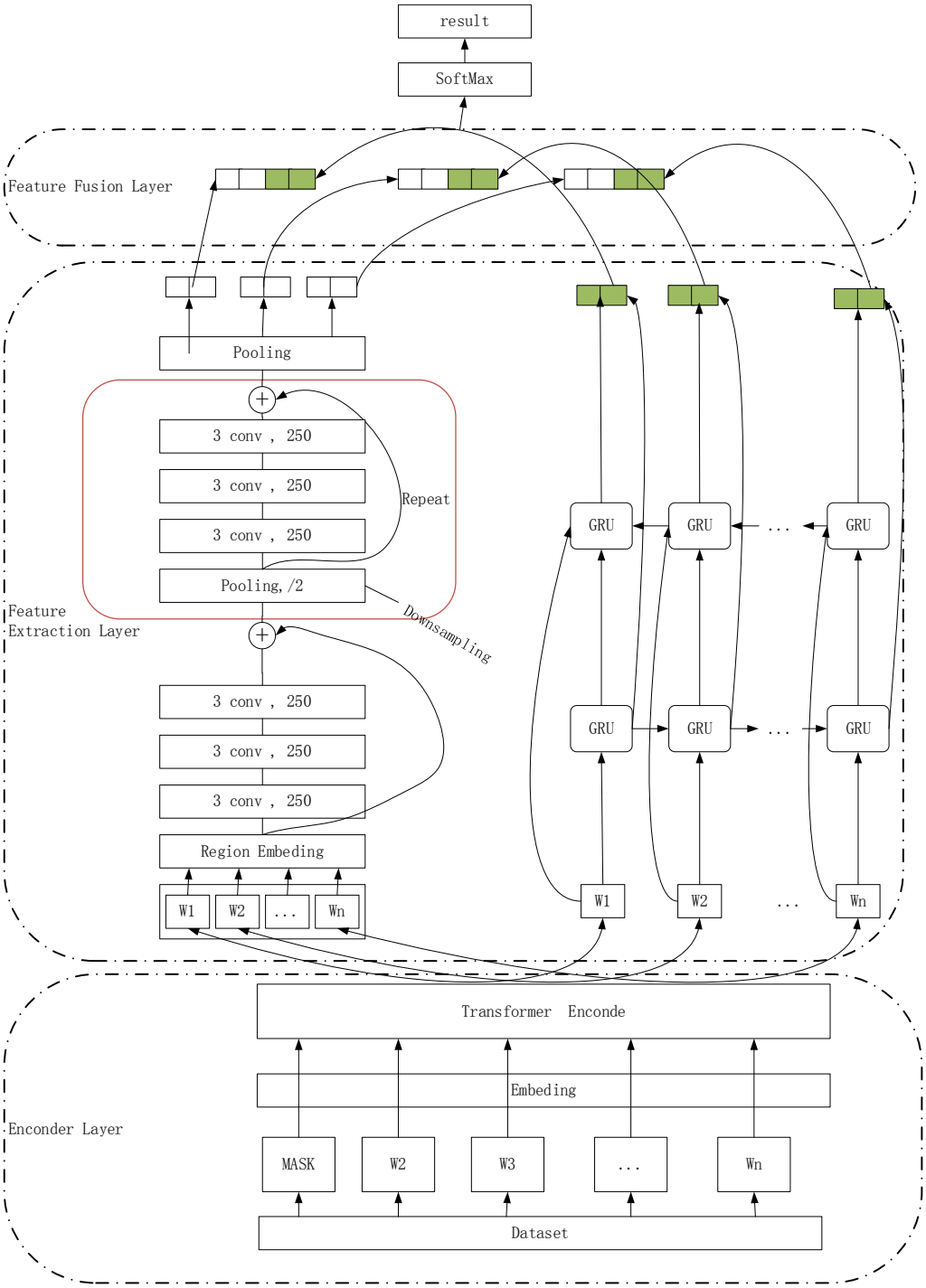


Figure 2. Model framework diagram.

In this paper, after considering the existing text classification models and the characteristics of the enterprise support policy text, we propose the BGC model, and the model framework is shown in Figure 2, which can be divided into six parts.

①The collected data are cleaned and segmented, and then converted into word vectors and input into the BERT model. The output of this model is a vector representation of each word that incorporates full-text semantic information, which is then passed to the DPCNN layer.

②The DPCNN layer uses two layers of equal-length convolution in order to improve the richness of the vector representation, which enables the model to learn more local features.

③The pooling layer selects steps 2 and 3 for the maximum pooling operation, which mainly serves to compress the length of the overall sequence, thus facilitating the model to capture long-range dependencies.

④And then the data is fed to the BiGRU layer to further extract textual contextual features to obtain richer and more complete feature information.

⑤Finally, the softmax layer is used to classify the data coming from the BiGRU layer. The dimensionality of each vector is obtained as 4 dimensions. The final category is determined by comparing the magnitude of the median value of each dimension.

⑥In order to solve the problems of gradient dispersion, data imbalance and hard samples, the model is improved: the gradient dispersion is introduced into the residual network, which makes the training of deep networks possible; the data imbalance and hard samples are introduced into FocalLoss, and the problem is also properly solved.

5. Experimental design

5.1. Data acquisition and pre-processing

In order to ensure the authority and accuracy of the data source, this system adopts the China SME Policy Information Dissemination Platform and the official website of the Chinese government as the data source. In order to save the trouble of subsequent data cleaning, this paper sets up a filtering method in the crawling stage, discarding all unnecessary elements except the document content itself to ensure the quality of the data source, and finally a total of 102,268 pieces of data are collected. Through pre-processing, this paper retains the content of the body part of the policy to facilitate the subsequent processing. The focus of this paper is to classify the text from the enterprise perspective, and also to deal with the problem of unbalanced categories and difficult samples, and to manually label the enterprise support policies using the four categories shown in Table 1, and the distribution of policy text classification data is shown in Table 2.

Table 2. Distribution of text data for 4 types of policies.

Policy text category	Quantity /period
Unrelated information	57781
Preferential Policies	8145
Reward Policies	31049
Individual related policies	5293

5.2. Text segmentation and type analysis

Text segmentation: By looking at the text content of the policy in the dataset, it can be intuitively seen that the key part of the policy text is generally composed of several clauses,

and each clause describes a wide range of information, not entirely for a certain type of enterprise, so this paper uses a rule-based text segmentation method to segment the text content.

Analysis of enterprise support policy types: After the text segmentation of the policy body content into several clauses, it is found that there is redundant information in these clauses for enterprises, which increases the difficulty of finding opportunities for enterprises, so this paper identifies the following four categories from the perspective of enterprises: irrelevant information, preferential policies, incentive policies, and personal related policies.

5.3. Categorized experimental setup

In this paper, we use Pytorch deep learning framework, GeForce RTX 3060 GPU, Python 3.7, Torch 1.11.0, and BERT model to achieve the task of multi-categorization of enterprise support policy text. Firstly, this paper performs category labeling according to the pre-defined categories, and then divides the enterprise support policy text dataset into training set, validation set and test set according to the ratio of 8:1:1. Finally, the pre-training model BERT-Base-Chinese is loaded, and the results are input into the BGC model, and the whole model experimental parameters are set as shown in Table 3.

Table 3. Experiment parameter settings.

Parameters	Parameter explanation	Value
num_epochs	Number of training rounds	5
batch_size	Number of samples per batch during training	8
pad_size	Length of each sentence processed to	256
learning_rate	Learning Rate	$5x10^{-5}$
require_improvement	Early end of training batch count	1000

In addition, in order to verify the effectiveness of the BGC model and the effect of this model on corporate support policy text classification, this paper conducts comparison experiments with the traditional BERT model. All the collected corporate-support policy texts are used as corpus, and the trained word vector model is loaded on the training texts for vector representation of corporate-support policy texts, which are input into the above 2 models for classification experiments, and the number of training samples and maximum sentence length of each batch are treated the same as the BERT model, with the parameters set as shown in Table 3, and the activation function is softmax, and finally the classification results of different Finally, the classification results of different models are compared and analyzed.

6. Experimental results and comparative analysis

For the effectiveness of business support policy text classification four indexes were used to evaluate the accuracy, precision, recall and F1 value, and the experiments were conducted under the same conditions, and the final experimental results are shown in Table 4.

Table 4. Experimental results.

Model	Accuracy	Precision
BERT	0.9525	0.9159
BGC	0.9729	0.9712

From the experimental results in Table 4, it can be seen that the BGC-based model has the best results in classifying business support policy text, and all categories are evaluated better than other models. When the α of FocalLoss is set to [0.4, 0.2, 0.2, 0.2] and γ is set to 2 for training, the accuracy, accuracy, completeness and F1 values can reach over 97%. From Figure 3, we can see that the classification of "irrelevant information" is the best, and its F1 value is more than 97%; "reward policy" and "preferential policy" are the next best, and the classification of personal-related policies is poor. The classification of policies is less effective. In general, the BGC model can achieve the automatic classification of enterprise support policy texts more accurately.

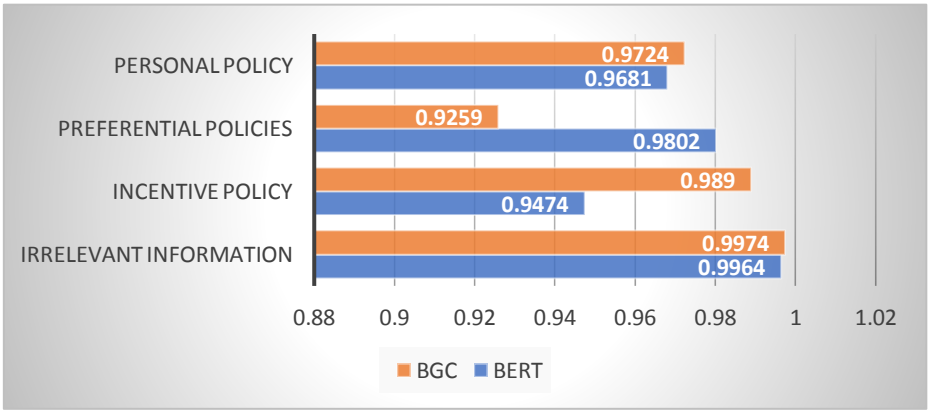


Figure 3. Categorical effect by policy category.

7. Conclusion

Driven by the construction of digital government, the research on policy texts in academia has just become a general direction. Whether the subsequent query or matching of policies is carried out, the first step must be to achieve automatic classification of policy texts, and this paper provides a reference example for text classification in the field of business support policies. From the perspective of small and micro enterprises, this paper breaks the limitation of the original policy classification model that considers the text as a whole, and proposes an automatic classification method for the text content of the text of business support policies based on the BGC model, and draws the following conclusions: first, the superiority of the BGC model in the field of business support policy text classification is verified in the comparison experiments of BERT and BGC models; then, by changing the Then, by changing the original cross-entropy loss function to FocalLoss, the classification effect is effectively improved, and it is confirmed that the loss function can not only solve the sample imbalance problem, but also be very effective for the hard sample problem; finally, the method achieves a more accurate automatic classification of corporate support policy text,

and the accuracy and F1 values of the best experiments reach over 90%. However, this paper also has many shortcomings: it only starts from the body part of the policy, and does not involve policy-related information such as the title.

The future application of automatic policy text classification is very broad, and there are still many follow-up works based on this model: firstly, the algorithm model is continuously improved while expanding the data set; secondly, the classification results are correlated and analyzed to obtain useful information; finally, the whole model is applied to policy recommendation to recommend suitable policies for enterprises.

Acknowledgments

This work was supported by the Shandong Province Soft Science Major Project "Research on the Construction of Science and Technology Financial System and Innovative Methods in Shandong Province" No.2022RZA03016 and Shandong Digital Convergence Publishing Innovation and Entrepreneurship Community No.2023JBGS05.

Conflicts of interests

The authors declare no conflict of interests.

References

[1] Mohajan HK. Cost Minimization Analysis of a Running Firm with Economic Policy. *Ann. Spiru Haret Univ. Econ. Ser.* 2022, 22(3), 317–337.
[2] The China issued the "Guidance on Strengthening Digital Government Construction. Available: https://www.gov.cn/xinwen/2022-06/23/content_5697326.htm (accessed on 23 June 2023).
[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* 2018, arXiv:1810.04805
[4] Hassan SU, Ahamed J, Ahmad K. Analytics of machine learning-based algorithms for text classification. *Sustain. Oper. and Comput.* 2022, 3:238–248.
[5] Gasparetto A, Marcuzzo M, Zangari A. A survey on text classification algorithms: From text to predictions. *Information* 2022, 15(8):46.
[6] Karaman Y, Akdeniz F, Savaş B K, *et al.* A Comparative Analysis of SVM, LSTM and CNN-RNN Models for the BBC News Classification. In *the Proceedings of the International Conference on Smart City Applications*. Cham: Springer International Publishing, 2022, 6, pp. 473–483
[7] Khan L, Amjad A, Ashraf N, Chang HT. Multi-class sentiment analysis of urdu text using multilingual BERT. *Sci. Rep.* 2022, 12(1):5436.
[8] Li Q, Xiao Z, Zhao Y. Research on the Classification of New Energy Industry Policy Texts Based on BERT Model. *Sustainability* 2023, 15(14):11186.
[9] Hagen L, Uzuner Ö, Kotfila C, *et al.*: Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *2015 48th Hawaii International Conference on System Sciences*, Kauai, HI, USA, 5-8 January 2015, pp. 2134–2143.
[10] Shim J, Park C, Wilding M. Identifying policy frames through semantic network analysis: an examination of nuclear energy policy across six countries. *Policy Sci.* 2015, 48:51–83.

[11] Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, *et al.* Semantics-aware BERT for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 9628–9635.
[12] Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, Jmaiel M. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Appl. Sci.* 2022, 12(6):2891.
[13] Zhao DD, Huang DG, Meng JN. Chinese entity relationship classification based on BERT-GRU-ATT model. *Comput. Sci.* 2022, 49(06):319–325.
[14] Zhao J, Guo S, Mu D. DouBiGRU-A: software defect detection algorithm based on attention mechanism and double BiGRU. *Comput. Secur.* 2021, 111:102459.
[15] Zhu Z, Dai W, Hu Y, Li J. Speech emotion recognition model based on Bi-GRU and Focal Loss. *Pattern Recognit. Letts.* 2020, 140:358–365.