sciendo

DOI: 10.2478/eces-2023-0008 ECOL CHEM ENG S. 2023;30(1):103-110

Ting ZHANG[1,2]

THE IMPACT OF GREEN FINANCE DEVELOPMENT ON ECOLOGICAL PROTECTION BASED ON MACHINE LEARNING

Abstract: In the context of today's green development, it is the core task of the financial sector at all levels to enhance the utilisation of resources and to guide the high-quality development of industries, especially to channel funds originally gathered in high-pollution and energy-intensive industries to sectors with green and high-technology, to achieve the harmonious development of the economy and the resources and environment. This paper proposes a green financial text classification model based on machine learning. The model consists of four modules: the input module, the data analysis module, the data category module, and the classification module. Among them, the data analysis module and the data category module extract the data information of the input information and the green financial category information respectively, and the two types of information are finally fused by the attention mechanism to achieve the classification of green financial data in financial data. Extensive experiments are conducted on financial text datasets collected from the Internet to demonstrate the superiority of the proposed green financial text classification method.

Keywords: green finance, text classification, attention mechanisms, machine learning

Introduction

With the growing problem of environmental pollution and the gradual depletion of resources for production, directing the focus of economic development towards green high-tech industries is an inevitable choice for sustainable economic development [1]. In this process, the government's various financial departments at all levels play the role of the helmsman, and they need to formulate reasonable policies according to the actual situation of the market, so that the capital originally gathered in the high-pollution, energy-intensive sunset industries will flow into the industries with green and high-tech. Before formulating these financial policies, staff from various government financial departments at all levels need to analyse the vast amount of textual data containing various financial sectors to filter out the green high-tech industries with potential for development. However, it is a complex and time-consuming task to sift through such a large amount of financial text data to identify green and high-tech industries with potential for development, and it is difficult to guarantee that there will be 100 % error-free sifting of financial data over a long period, which inevitably affects the scientific and rational nature of the final policy. Therefore, it is of great value to study how to make use of the rapidly developing natural language processing technology to manage financial text data efficiently and to

[1] Zhongyuan Institute of Science and Technology, Zhengzhou 450000, China, ORCID: 0009-0008-6498-8229
[2] Henan Pivotal Port and Economy Research Centre, Zhengzhou 450000, China, email: tztingzhang@163.com

provide government staff at all levels and in all financial departments with the accurate content of green financial texts from the huge amount of financial data.

Natural language processing is the use of computers to analyse incoming linguistic data and mine it for useful text data for human-computer interaction. One of the basic tasks is financial text classification, where pre-tagged financial text data is trained to automatically classify the required green financial text data from a large amount of financial text. The initial text classification task was performed manually, which is a simple method but requires a high level of expertise in the text to be classified and limited accuracy in processing large amounts of data manually. Later, due to the rapid development of machine learning, there emerged decision tree models such as classification by restricting rules [2], plain Bayesian [3] classification methods that obtain classification results by finding probabilities, support vector machines [4], and K-nearest neighbour models [5] that calculate distances between data by statistical means.

Recent research using deep learning on financial text data has avoided the various problems that exist in machine learning. Networks such as Convolutional Neural Network (CNN) [6], Recurrent Neural Network (RNN) [7], and attention mechanisms [8] in deep learning avoid the complex manual work of extracting features and greatly improve the performance of financial text classification by automatically learning features.

To further improve the accuracy of green financial text classification in financial text data, and to ensure that governments and financial institutions at all levels can formulate appropriate economic policies based on the classification results so that the economy can develop in a green and healthy way. This paper proposes a parallel BERT green financial text classification model with an attention mechanism, which can well identify the green financial text data in the financial text data.

Related work

The classification of financial text data containing green financial text data is an important area of research. In the government financial sector, deep learning techniques can help staff to filter green financial text data from financial data and reduce the workload of staff in the face of large amounts of financial data, and more and more researchers are using natural language processing techniques to classify financial text data.

With the widespread use of CNNs in computer vision, many researchers have started to try to apply CNN methods to text classification. TextCNN [9] model improves the original CNN input layer suitable for image data to one suitable for text data, but the convolutional kernel size in TextCNN is fixed, which can only extract local features in text of fixed length, and cannot focus on the key information in text with a longer interval. The key information at longer intervals cannot be focused on. For the model to focus on the global information of the text, the long short term memory (LSTM) [10] mimics the human state when remembering, using three gates with a special structure that allows the network to control the information flowing through, improving the classifier's ability to capture the global information. Such methods show excellent classification performance when the text length is short, however, if the length of the text data is very long, the spacing of key information in the text data will be large, weakening the LSTM's memory for the key information and causing the LSTM's performance to drop sharply when faced with long text. Therefore, some researchers have considered combining the CNN's ability to extract

local information from text with the LSTM's ability to extract global temporal information to improve the model's performance in the task of classifying long English papers [11].

However, both the CNN model, which extracts local information, and the LSTM model, which collects global temporal information, do not take into account the overall structured features of long text, and there is still room to explore the performance of text classification. At the same time, the input data of the above models are all obtained by direct coding, which poses a significant burden in terms of hardware overhead. The proposed attention mechanism allows the model to focus on meaningful words in the text data to extract information from them that is important for the classification task, and the Self-Attention mechanism [12] is used to achieve parallelised and accelerated training. BERT [13] was inspired by Bi-LSTM [14] and uses a bi-directional Transformer to extract key information in the network. Some researchers have proposed hierarchical attentional hybrid sparse networks in long text classification models [15]. Combining RNN with a self-attention mechanism applied at the word and sentence level, not only does the long text structure as a text semantic feature incorporated into the classification model but also the RNN model used is sparse and the self-attention mechanism is used to focus on the key words and sentences to build the text classifier model [16].

Green finance text classification model

The double BERT green financial text classification model proposed in this paper is mainly composed of input module, data analysis module, data category module and classification module. After inputting green financial texts to the input module, the data features of green financial texts are obtained in the data analysis module, the category features in green financial texts are extracted in the data category module, and then the two features are input to the classification module, and finally the categories of input green financial texts are predicted by linking the data features with the categories. The data analysis module and the data category module in the proposed model use the same structure to extract different features, and both modules contain a BERT-based word embedding layer and coding layer. The classification module, on the other hand, contains a transformation layer, a fusion layer, and a classification layer.

Input module

For both the data analysis module and the data category module, the input sequence entered by the input module is a sequence of sentences in a batch of data set denoted as S^N, and N denotes the batch size. One of the sentences can be denoted as:

$$S_j^N = \left(w_{j1}, w_{j2}, \ldots, w_{ji}, \ldots, w_{jn}\right)$$

where j is the j-th sentence in this batch of green finance text sentences, w_{ji} is the i-th word in the j-th sentence, and n is the sentence length.

Word embedding layer

To obtain data features and category features in green finance text, Our dual BERT uses a BERT pre-training model to map words in the text from a low-dimensional space to a multi-dimensional space one by one, providing a larger model of language understanding that can yield better information about the data. In this paper, the two BERT pre-training models are trained with the category features and data features of the green finance text

respectively, so that the model can fully obtain the various word embeddings in the green finance text. The input of the BERT pre-training model is the original word vector of the green finance text S after the random initialisation and word separation process.

Coding layer

The task of the coding layer is to encode the embedding vector transformed by the word embedding layer into a sequence vector containing rich contextual semantic information. The encoding layer uses a three-layer neural network to encode the word embedding vector. The input dimension of the input layer is the embedding dimension of the word embedding layer, the hidden layer dimension is 512, and the output dimension of the output layer is 3. The training objectives of the data category module and the data analysis module are different, but the others are the same. The training objectives of the data category module are 0, 1(p,q), 2(p,q), p, q are two entities in the green financial text, and the training objectives of the data analysis module are 0, 1, and 2. The three-layer neural network uses the ReLU activation function to enhance the nonlinearity of the neural network model, and the ReLU activation function can accelerate the training speed compared with other activation functions.

Conversion layer

The conversion layer mainly uses a linear transformation to calculate the variable length of data features and category features.

If h_{t-1} and b_{t-1} represent the data features and category features before entering the conversion layer, W_1 and W_2 represent the conversion parameters, b_1 and b_2 represent the bias, and h_t and b_t represent the data features and category features after conversion, then the calculation formula of the conversion layer is:

$$h_t = W_1 h_{t-1} + b_1 \tag{1}$$

$$b_t = W_2 b_{t-1} + b_2 \tag{2}$$

Fusion layer

The task of this layer is to fuse the data features and category features of the same length obtained from the transformation layer to prepare for the classification in the classification layer. This model mainly uses the attention mechanism to fuse the data feature vector h and the category feature vector b of the text and calculates the attention score a for both. The attention mechanism originates from the fact that humans selectively focus on the key parts of all information while ignoring other visible information. The attention mechanism used in our proposed double BERT is a modified scaled feature attention.

The attention mechanism is calculated by first multiplying the data feature h and the category feature b, then dividing the length of the data feature h and the category feature b by the scaling operation, and the result can be chosen whether to perform the mask operation or not and then performing the softmax calculation. Finally, the attention score a of data feature h and category feature b is obtained by dotting the result with category feature b. This process is calculated by the following formula:

$$a(h,b) = soft\max\left(\frac{hb}{\sqrt{d}}\right)b \tag{3}$$

The model treats the attention score as the final feature link between the different categories of features identified in the green finance text domain, maps it to the resulting space required by the task through a linear network, and uses softmax to calculate the probability that the relationship between the different data categories is *y*:

$$y = soft\max\left(W_a\left(h,b\right)+l\right) \tag{4}$$

where *W* is the matrix that maps the vector to the output vacuum; *l* is the bias.

The *LOSS* used by the model is:

$$LOSS = -\left(g_i, \ln y + (1-g_i)\ln(1-y)\right) \tag{5}$$

where g_i is the vector representation of the labels; *y* is the true sample label.

Experiment

Experimental hardware: CPU is i7-10700k, memory capacity is 16 GB, the graphics card is RTX2080, video memory capacity is 8 GB, TensorFlow2.0 deep learning framework.

To verify the validity of the proposed model, a green finance dataset containing 361,744 green finance text data was crawled from the Wind financial database in this paper, and the specific composition is shown in Table 1.

Table 1

Green finance text dataset

Green Credit	Major banking institutions green credit situation table	2013-2019
Green Credit	Breakdown of green credit investment by major banking institutions	2013-2019
Green Credit	Green Credit Fact Sheet (by Bank) [year]	2008-2019
Green Credit	Green Credit Support Project Environmental Benefits Table (by Bank) [year]	2007-2019
Green Credit	Details of green credit investment (by Bank) [year]	2008-2019
Green Credit	Breakdown of loans to the two high and one leftover industries (by Bank) [year]	2007-2019
Green Credit	Green Credit Card Business Statistics Table (by Bank) [year]	2009-2019
Green Bonds	Table of information on green bonds issued domestically [day]	2016-2019
Green Bonds	Information Table of Green Bonds Issued Overseas by Domestic Entities [day]	2015-2019
Green Operation	Table of Green Affair Operations (by Bank) [year]	2005-2019
Carbon Finance	China Carbon Emissions Trading Information Sheet [day]	2013-2019

Comparison experiment

In order to verify the superiority of the model in text classification, TextCNN, BERT, BERT+RNN, BERT+CNN, BERT+Bi-LSTM, BERT+Transformer and models in this paper are compared for different evaluation metrics. While comparing the classification ability of different models on the green finance text dataset crawled in this paper, three Chinese datasets RCV1-V2, simplify _ 4_moods (abbreviated as s4m), online_shopping_ 10_cats (referred to as os10c) to test the performance of the model in handling different text classification tasks [14]. To prevent chance results in a single experiment, the various models were run 10 times to find the mean values, and the classification results obtained on

Accuracy, Precision, Recall, and F1 measure value evaluation metrics are shown in Tables 2-5.

Table 2

Classification accuracy

Network name	RCV1-V2	s4m	os10c	ours
TextCNN	0.63	0.61	0.56	0.58
BERT	0.70	0.63	0.58	0.60
BERT+RNN	0.82	0.73	0.72	0.71
BERT+CNN	0.86	0.83	0.86	0.78
BERT+BiLSTM	0.87	0.83	0.87	0.79
BERT+Transformer	0.89	0.87	0.83	0.83
Ours moudle	0.89	0.92	0.91	0.96

From the comparison of Accuracy rates in the Table 2, the TextCNN model has the lowest accuracy on each dataset, followed by BERT. This is because these two methods cannot fully extract various features in the data. The downstream task can capture some structural information by CNN or BiLST, but the extraction ability is limited. The accuracy of our model is the highest in all datasets, and it is on par with BERT+CNN, BERT+BiLSTM, and BERT+Transformer models in the RCV1-V2 dataset, and reaches 0.92 in the s4m dataset, which is higher than the best BERT+Transfer model (0.87) in the os10 dataset. The accuracy on the os10c dataset is greater than that of the best BERT+BiLSTM model (0.87), and the accuracy on the s4m dataset is 0.92, which is higher than that of the best BERT+Transformer model (0.87), and the accuracy on the os10c dataset is 0.96, which is much higher than that of the other models.

Table 3

Classification precision

Network name	RCV1-V2	s4m	os10c	ours
TextCNN	0.65	0.65	0.62	0.54
BERT	0.75	0.67	0.60	0.64
BERT+RNN	0.80	0.79	0.71	0.69
BERT+CNN	0.83	0.89	0.85	0.75
BERT+BiLSTM	0.86	0.83	0.83	0.77
BERT+Transformer	0.88	0.86	0.83	0.81
Ours moudle	0.88	0.93	0.90	0.95

Table 4

Classification recall

Network name	RCV1-V2	s4m	os10c	ours
TextCNN	0.72	0.63	0.69	0.63
BERT	0.75	0.61	0.61	0.70
BERT+RNN	0.83	0.76	0.75	0.75
BERT+CNN	0.89	0.82	0.85	0.81
BERT+BiLSTM	0.90	0.86	0.83	0.82
BERT+Transformer	0.86	0.87	0.82	0.83
Ours moudle	0.90	0.91	0.89	0.94

From the comparison of Precision rates in the above table, the proposed model gets the same Precision on the RCV1-V2 dataset as the BERT+Transformer model, slightly higher

than the BERT+CNN and BERT+BiLSTM models, and higher than the TextCNN, BERT model, and BERT+RNN model. The highest Precision is obtained on the datasets s4m, os10c, and this paper, which are higher than the best 0.86, 0.83, and 0.81 of the comparison models, respectively, reflecting the effectiveness of this paper's model for the text classification problem.

From the comparison of Recall in the above table, the model in this paper achieves optimality on all datasets and is on par with the BERT+BiLSTM model on the RCV1-V2 dataset. In the s4m dataset, the recall of this model and the BERT+Transformer model is 0.91 and 0.87 respectively, which is better than the comparative experimental model. In the os10c dataset, the recall of this model is 0.89, which is higher than that of the best BERT+CNN model (0.85), and in the green finance dataset, the recall of this model is 0.94, which is much higher than that of the best BERT+Transformer model (0.83). The model has been effectively improved.

Table 5

Classification F1 measure

Network name	RCV1-V2	s4m	os10c	ours
TextCNN	0.68	0.64	0.65	0.67
BERT	0.75	0.64	0.60	0.72
BERT+RNN	0.81	0.77	0.73	0.77
BERT+CNN	0.86	0.87	0.85	0.81
BERT+BiLSTM	0.88	0.84	0.83	0.80
BERT+Transformer	0.87	0.86	0.85	0.82
Ours moudle	0.89	0.92	0.90	0.93

From the comparison of F1 measure in the above table, this model achieves the same value as the BERT+BiLSTM model on RCV1-V2 dataset, 0.92 on the s4m dataset, 0.87 and 0.86 for the BERT+CNN model, and BERT+Transformer model respectively, TextCNN model and BERT model have the lowest value. The F1 measure of this model reaches 0.90 in the os10c dataset, which is higher than the best BERT+Transfer model (0.85), and the highest F1 value of this model reaches 0.93 in the green finance dataset, which is higher than all the models compared.

The above comparative experiments show that the proposed dual BERT Green Financial text classification model based on the attention mechanism performs well in all aspects. Compared with the comparative experimental model, the classification performance has been improved to varying degrees, and it can complete the accurate task of green financial text data classification.

Conclusion

In this paper, a dual BERT green finance text classification model with an attention mechanism is designed. The overall design of the model includes an input module, a data analysis module, a data classification module, and a classification module. The experimental results show that the classification accuracy of the model proposed in this paper reaches 96 % for the collected green finance data, which can effectively improve the efficiency of personnel at all levels in the green finance sector and achieve the coordinated development of the economy, resources, and environment.

Acknowledgements

This work was supported by Henan Philosophy and Social Science Planning Project "Study on the measurement and strategy of Science and Technology - Ecology - Economy Coupling coordinated development at county level in Henan Province", the project number is 2022BJJ118.

References

[1] Wang YX. Research on influence of computer technology on industrial upgrading and cooperative evolution of environmental protection fiscal policy from green finance perspective. J Physics: Conf Ser. 2021;1744(4). DOI: 10.1088/1742-6596/1744/4/042104.

[2] Han ZQ, Wen LN. Development and validation of a decision tree classification model for the essential hypertension based on serum protein biomarkers. Annals Translat Medicine. 2022;10(18):1-6. DOI: 10.21037/ATM-22-3901.

[3] Wang HM, Li ZJ. An AdaBoost-based tree augmented naive Bayesian classifier for transient stability assessment of power systems. Proc Institution Mechanical Engineers. 2022;236(3):495–507. DOI: 10.1177/1748006X211047308.

[4] Tang JY, Lin KY, Li L. Using domain adaptation for incremental SVM classification of drift data. Mathematics. 2022;10(19):3579. DOI: 10.3390/MATH10193579.

[5] Kim YK, Kim HJ, Lee H, Chang JW. Privacy-preserving parallel kNN classification algorithm using index-based filtering in cloud computing. PLoS ONE. 2022;17(5):e0274981. DOI: 10.1371/JOURNAL.PONE.0267908.

[6] Lewy D, Mańdziuk J. Training CNN classifiers solely on webly data. J Artificial Intelligence Soft Computing Res. 2023;13(1):75-92. DOI: 10.2478/JAISCR-2023-0005.

[7] David MS, Renjith S. Comparison of word embeddings in text classification based on RNN and CNN. IOP Conf Ser: Materials Sci Eng. 2021;1187(1):012029. DOI: 10.1088/1757-899X/1187/1/012029.

[8] Mundra S, Mittal N. FA-Net: fused attention-based network for Hindi English code-mixed offensive text classification. Social Network Analysis Mining. 2022;12(1):1-6. DOI: 10.1007/S13278-022-00929-1.

[9] Rawat A, Wani MA, ElAffendi M, Imran AS, Kastrati Z, Daudpota SM. Drug adverse event detection using text-based convolutional neural networks (TextCNN) technique. Electronics. 2022;11(20):3336. DOI: 10.3390/ELECTRONICS11203336.

[10] Ji XD, Li YL, Xia Y, Wang CW, Ansari F, Wu B, et al. LSTM approach for condition assessment of suspension bridges based on time-series deflection and temperature data. Adv Structural Eng. 2022;25(16):1-10. DOI: 10.1177/13694332221133604.

[11] Abdul AI, Marina Y. An efficient hybrid LSTM-CNN and CNN-LSTM with GloVe for text multi-class sentiment classification in gender violence. Int J Adv Computer Sci Appl (IJACSA). 2022;13(9):1-11. DOI: 10.14569/IJACSA.2022.0130999.

[12] Bae A, Kim W. Speaker verification employing combinations of self-attention mechanisms. Electronics. 2020;9(12):2201. DOI: 10.3390/ELECTRONICS9122201.

[13] Huang WC, Tao ZQ, Huang XH, Xiong LY, Yu J. Hierarchical self-attention hybrid sparse networks for document classification. Mathematical Problems Eng. 2021:1-10. DOI: 10.1155/2021/5594895.

[14] Peng HL, Tsu JF, Wei YM. Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. Proc AAAI Conf Artificial Intelligence. 2020;34(05):8236-44. DOI: 10.1609/aaai.v34i05.6338.

[15] Wu CH. An empirical study on discussion and evaluation of green university. Ecol Chem Eng S. 2021;28(1):75-87. DOI: 10.2478/eces-2021-0007.

[16] Wu CH, Tsai SB, Liu W, Shao XF, Sun R, Wacławek M. Eco-technology and eco-innovation for green sustainable growth. Ecol Chem Eng S. 2021;28(1):7-10. DOI: 10.2478/eces-2021-0001.