Application of BERT Model in Forecasting Introduction Text and Concept Section of Listed Companies

Boyu Pan[1, a, *]
[a] e-mail: 2130102050@st.btbu.edu.cn

[1] School of International Economics and Management, Beijing Technology and Business University, Beijing, China

Abstract: The stock plate refers to the fact that certain companies have certain specific relevant elements in the stock market, which is named after this element. Plate classification are mainly industry classification and concept classification. The China Securities Regulatory Commission has a unified classification standard for the industry sector, but there is no such standard for the conceptual sector. A stock often involves many conceptual plates. Understanding the conceptual plate classification of stocks can be a very basic reference condition for investors. Therefore, this paper attempts to analyze the description text of listed companies through the BERT model to complete the multi-label classification task of their respective plates, and informationize a large number of texts into simple and clear label information, which can be used as a reference for further analysis of the stock information.

Keywords: listed company, sector, multi-label classification, BERT

1. Introduction

As one of the important technical fields of artificial intelligence, natural language processing technology has made great progress in recent years. In the era of financial innovation, NLP technology has been used to deal with unstructured information in the financial field, including a large amount of text information. This paper aims to save users ' time and energy in preliminary screening in the face of a large number of text content by processing a large amount of text information and pre-processing as a key element of conciseness.

The concept section of listed companies is a major reference for investors to screen investment stocks, and its popularity in the market continues to rise. The concept section can be said to be the induction and understanding of this series of stocks with the same attributes after being judged by the market. Investors will select stocks based on the active concept sections mentioned in the market, and a stock can involve many concept sections at the same time. And many times these sections are not the main business scope of the listed company. Therefore, if you want to select stocks involving the target section, you need to do some information screening work.

Through the analysis of the text information of listed companies obtained at present, this paper believes that this goal can be achieved through the multi-label classification task, and through the text description analysis of the listed company itself and its operating projects to learn the concept plate situation.

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2332273

2. Research status

There are many tool models in the field of natural language processing analysis, and there are many alternative implementation models in text multi-label classification. Text multi-label classification was originally decomposed into multiple independent binary classification problems (Boutella et al.) [1], It is equivalent to training a binary classifier independently for each label. Later, Kim et al. first used the CNN structure for text classification [2], using multiple convolution kernels to extract text features and then output the probability distribution of the label space. Lai[3] combined the advantages of RNN and CNN to propose the RCNN model, aiming at the problem that traditional classification methods ignore context and the problem of CNN convolution window setting.

One of the most popular models in recent years is the BERT model proposed by Jacob Devlin et al. [4], which obtains features containing text semantics by training large-scale unlabeled corpus. Based on BERT, researchers can achieve different goals by designing different downstream tasks, one of which is multi-label classification of text. CorNetBertXML (Xun et al., 2020) [5] used BERT to obtain a joint representation of text and all candidate labels, and used additional exponential linear units (ELU) in the prediction layer to utilize label-related knowledge. Qian-Wen Zhang et al. [6]established a joint embedding mechanism of label and text, and learned the label itself as part of the input. On this basis, Ximing Zhang [7] et al. further established a multi-task learning multi-label classification model, and used the joint embedding mechanism to simultaneously obtain the correlation between text and label by learning the correlation between labels to improve the effect of multi-label classification.

3. Experimental methods

3.1 BERT multi-label classification model

The BERT model is a word vector that uses Transformer to generate context. It can learn complex semantic representations from the text. Users can directly inherit the pre-trained language model, and further fine-tune the model to make the model more suitable for the needs of users. Strong generalization ability has excellent performance in various tasks such as topic classification, question answering tasks, etc. In some recent studies, BERT has also shown its strong superior performance in text multi-label tasks.

In multi-label tasks, many studies tend to embed labels and texts as a joint, based on Transformer encoders to obtain a joint representation of the two. In addition, many researchers have begun to pay attention to the correlation between labels. Tsoumakas and others believe that a set of labels is semantically related to the text content, and the correlation between labels should be fully utilized to establish a multi-label classification model with generalization ability. Vinyals et al.captured the correlation between tags based on the seq2seq method, but found that the method relied heavily on the order relationship between tags. Zhang et al.referenced the shared encoder based on the BERT model, and constructed two auxiliary tasks for learning to predict the co-occurrence relationship between labels and the correlation between labels. The model strengthens the learning of label correlations and ultimately has a better performance in multi-label prediction classification.

In this paper, we use the auxiliary task to learn the multi-label classification model. The text and label are input into the BERT model as joint input, and the text and label representations are obtained at the same time. Based on the shared encoder, a label co-occurrence task is constructed outside the multi-label classification task to assist in improving the effect of the multi-label classification task.

3.2 Experimental conditions

(1) Text material

This article obtains a total of 1275 introduction texts about companies from 155 listed companies and the concept section labels involved in these 155 listed companies from the relevant webpages of Oriental Wealth. First of all, all the concept sections are for the entire listed company 's stock, and multiple introduction texts of the same listed company correspond to the same plate changes. Therefore, 1275 introduction texts are integrated through the company 's stock code and company name as conditions, and finally 155 long texts are obtained from the company.

In terms of labels, the concept section of the 155 listed companies is used as an input label. First, the labels of the entire 155 listed companies are screened and separated to establish a custom dictionary for the concept section of the data set. Based on the custom dictionary, word frequency statistics are performed on all concept block labels, and labels are matched with long text according to the company to construct a multi-label data set for this concept block prediction.

(2) Auxiliary task learning multi-label classification model

Analyze the important features of the current construction data set. First, after the company 's introduction text is integrated in the listed company as a unit, each text is very long, so it is necessary to select a model that can learn the complex semantics of long text. Secondly, as labels, there is a strong correlation between the concept blocks, and the co-occurrence between labels can be learned through auxiliary tasks, especially for low-frequency labels. Many scholars have proved that learning the correlation between labels can help improve the final multi-label classification results.

Combining the data features and model features that need to be analyzed, the multi-task learning multi-label model with label co-occurrence prediction as the auxiliary task is finally used as the basis to try to establish a prediction model suitable for generating the label of the concept plate according to the company 's introduction text. In addition to entering the multi-label classification module, the joint representation of the shared encoder is also used by the auxiliary task module, the label co-occurrence prediction module.

Based on the multi-task model, the integrated company introduction text and the concept plate label are used as common inputs, and the joint representation of the two is obtained through the BERT shared coding layer. In the multi-label classification task module, a multi-label text classifier is established through the fully connected layer, and the binary classification cross entropy is used as the loss function. In the auxiliary task tag co-occurrence module, the tag set Y is extracted by randomly extracting tags from the tag set corresponding to each document, and then the remaining tags in the overall tag set are predicted to be Y-related.

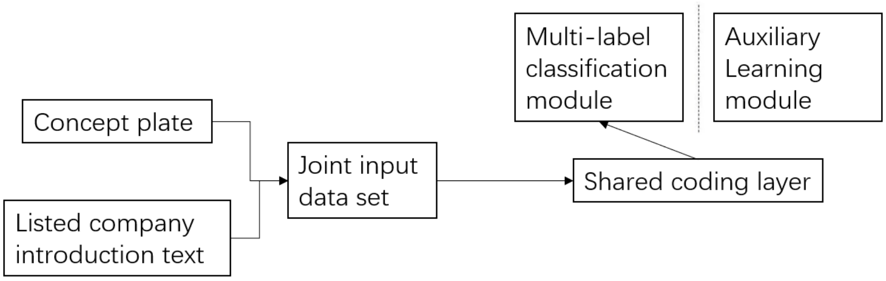
Concept plate
Listed company introduction text
Joint input data set
Shared coding layer
Multi-label classification module
Auxiliary Learning module

Figure 1 Concept Plate Prediction Process

(3) Model evaluation indicators

The prediction effect of the final concept plate label is determined by the effect of the multi-label classification task. According to the different emphases reflecting the task effect, hamming loss and Micro-F1 are used as the main indicators. According to the evaluation indexes of the researchers on the effect of multi-label classification model, Micro-P and Micro-R are used as auxiliary analysis indexes.

4. Experimental results

Referring to the previous experimental results of the researchers, the multi-label classification model based on BERT benefits from its own powerful text feature extraction ability. In comparison with the previous BR, Textcnn, seq2seq models, it can be seen from the main indicators HL and F1 indicators that the overall effect of the BERT model is more superior. Therefore, this paper mainly compares the BERT model with the joint embedding method to obtain the text-label joint representation and whether the model after adding the label correlation learning module can better adapt to the task of predicting the concept plate label.

We use the same listed company concept block dataset to test the multi-label classification model with auxiliary learning tasks and the multi-label classification model without auxiliary tasks, and also without document-label joint input. Both models use the BERT model as the baseline. The results are as follows:

Table 1: for different models on concept plate datasets.

Model	**HL**	**Mi-p/R/F1**
BERT+multi-label classification	0.028	49.2/36.0/39.2
Auxiliary tasks + multi-label classification	0.026	51.5/36.9/43.0

It can be seen that the addition of auxiliary task modules and the adoption of a joint embedded input data model have a certain degree of reduction in the hamming loss index, and the F1 index score is also higher, which proves that in the application of the prediction concept section, the method of learning the correlation between labels can indeed effectively improve the prediction effect.

On the other hand, compared to the BERT model 's performance on other publicly available large data sets, the concept block label prediction is indeed quite inferior. The reason is that in addition to training biases, there are several factors that have a great impact on the data set predicted by the concept block.

The first point is the size of the data set. The prepared data set has obtained thousands of text documents when collecting relevant data. However, since the concept plate label is for the listed company 's stock itself, the long text data after integration according to the company and stock code is limited by the number of companies, although the size remains unchanged. The BERT model 's performance in the fine-tuning phase can be quite different from large public data sets.

The second point is that the concept plate itself contains a wide range of plates. In addition to the plates such as domestic chips, 5G concepts, etc., which can be extracted in the company 's text introduction, such as sub-new shares, margin trading and other high-frequency concepts. Although the plate is the information that needs to be marked, it is often not reflected in the company introduction text that we can collect, which will affect the final prediction effect of the model. And because many or even high-frequency labels appear in such concept blocks, even if sub-modules that learn the correlation between labels are used, the label combination that the model can generate will still be affected.

Finally, about the convergence speed of the model, the BERT model used in this paper is composed of 12-layer encoders. The multi-label classification task entering the downstream through the full connection layer is itself a ' narrow and deep ' model. After integration, the text introduced by each company is extremely long, which will also affect the efficiency of model operation.

5. Conclusion

The multi-task multi-label classification model of the auxiliary learning task module that uses the joint representation of text and labels and adds the label co-occurrence relationship performs better than the multi-label classification model that only uses the BERT model plus the multi-label classification module on the task of predicting the concept block to which the listed company belongs through the introduction of text.

However, from the perspective of target realization, due to the limitations of the data set itself, its final effect still needs to be further improved. The data set needs the addition of appropriate data, further improves the label correlation learning module, and expands the auxiliary task module of the new label connection for the concept module label, thereby improving the final concept plate prediction effect.

Acknowledgments

National Natural Science Foundation of China [grant numbers 72171004] and the Humanity and Social Science Youth Foundation of Ministry of Education of China [grant numbers 21YJCZH186]

References

[1] Matthew R Boutella, Jiebo Luob, Xipeng Shena, and Christopher M Browna. (2004). Learning multi-label scene classification. Pattern Recognition, 37:1757– 1771

[2] Yoon Kim. Convolutional Neural Networks for Sentence Classification.[J]. CoRR, 2014, abs/1408.5882.

[3] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. In AAAI (Vol. 333, pp. 2267-2273).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages4171-4186

[5] Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. (2020). Correlation networks for extreme multi-label text classification. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1074–1082.

[6] Qian-Wen Zhang, Ximing Zhang, Zhao Yan, Ruifang Liu, Yunbo Cao, Min-Ling Zhang. (2021) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Main Track. Pages 3363-3369. https://doi.org/10.24963/ijcai.2021/463

[7] Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, Yunbo Cao, (2021), Computer Science, Computation and Language (cs.CL)

[8] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, pages 1171–1179.

[9] Syed Muzamil Hussain,and Chung SunTae(2021)."MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain." Applied Sciences 11.13. doi:10.3390/APP11136007.

[10] Menggang Li, et al. (2020)"Applying BERT to analyze investor sentiment in stock market." Neural Computing and Applications 33.10. doi:10.1007/s00521-020-05411-7.

[11] Wei Yang, Haotian Zhang,and Jimmy Lin. (2019)"Simple Applications of BERT for Ad Hoc Document Retrieval.." CoRR abs/1903.10972..

[12] Zhang Haifeng et al. (2022)' A news topic text classification method combining BERT and feature projection network. ' Computer application 42.04: 1116-1124.