Forecasting Financial Distress of Listed Firms Based on Recurrent Attention Networks

Ming Jia

jiaming9875@163.com

School of Computer Science and Engineering, Beijing Technology and Business University, Beijing, China

Abstract. Financial distress not only poses a threat to the long-term survival of a company, but also may have a chain reaction on the whole economic system. In recent years, the use of textual information as a feature of financial distress prediction has become a new hotspot, and this study proposes a financial text processing model based on recurrent attention network (RAN). Through an empirical study of A-share listed companies from 2007 to 2019, it is found that the RAN model performs well in extracting information from annual reports and effectively improves the accuracy of financial distress prediction.

Keywords: Self-attention; financial distress prediction; BERT; text processing

1 Introduction

With the rapid development of the global economy, listed companies play an increasingly important role in the market. However, the number of listed companies in trouble or even bankruptcy due to financial distress has been increasing in recent years, which has aroused great concern among investors, analysts and regulators. Financial distress not only poses a threat to the long-term survival of a company, but may also have a knock-on effect on the entire economic system. Therefore, accurate prediction of financial distress of listed companies has become an urgent problem.

Traditional forecasts of financial distress rely mainly on quantitative data such as financial ratios and ignore the large amount of qualitative information contained in the text. Annual report is an important document for listed companies to disclose their financial status and operating results to investors, creditors and other stakeholders. With the development of natural language processing and text mining techniques, scholars have begun to explore how to utilize these textual data to predict a company's financial distress. For example, Priyank Gandhi et al [1] studied sentiment indicators in bank annual reports and found that negative sentiment in annual reports is highly correlated with the probability of delisting; Mai et al. (2019) [2] processed management discussion and analysis through Word2Vec, transformed words into word vectors and average weighted them as textual features to predict bankruptcy of US companies; Jiang et al. (2021) [3] weighted BERT word vectors via tf-idf in order to make predictions about financial distress of Chinese unlisted companies.

ICBBEM 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347392

Although these studies of financial distress using the text of annual reports can yield good results, there are still some unresolved issues: for example, annual reports tend to be very long, and utilizing only word-level information is not sufficiently comprehensive and can result in the loss of sequential information. These problems lead to a reduction in the accuracy of forecasts. Therefore, how to extract the real valuable information from the text, how to utilize the text information efficiently and how to improve the accuracy of prediction are all important issues that we are facing.

The purpose of this paper is to address the above issues and to explore how this information can be efficiently utilized to improve the accuracy of financial distress forecasts. Specifically, the paper is structured as follows. Section 2 describes the research sample and how the text of the annual reports was processed; Section 3 describes the model proposed in this study; Section 4 conducts the experiments; and Section 5 presents the conclusions.

2 Sample and Text Processing

2.1 Sample

The sample of the study is A-share listed companies in Shanghai and Shenzhen from 2007 to 2019, which contains a total of 30,120 samples. According to the relevant regulations of capital market in China, if ST is added in front of the stock name of a listed company, the listed company is subject to risk warning, and the content of the warning contains various matters such as asset status, audit opinion, performance, frozen bank account, etc., which is comprehensive and covers a wide range of issues, therefore, in this paper, whether a listed company is ST is used to define whether it is in financial difficulties or not. Table 1 demonstrates the number of ST companies and non-ST companies in each year from 2007-2019.

Table 1. Number of ST and non-ST companies, 2007-2009.

years	ST	non-ST	years	ST	non-ST
2007	1084	123	2014	2420	40
2008	1085	108	2015	2503	36
2009	1345	116	2016	2768	55
2010	1644	126	2017	3181	55
2011	1911	125	2018	3264	65
2012	2126	89	2019	3445	113
2013	2246	47			

In this paper, we use a web crawler to obtain annual reports of listed companies from Oriental Wealth Online. However, annual reports contain a lot of useless and redundant information, and because Management Discussion and Analysis (MD&A) contains a lot of valid information, such as the company's management's analysis of the business, its outlook for the future, and the challenges it faces. This contains many details and clues about the inner workings of the company. We can extract valuable information from these texts and use it to predict the

financial status of the company. Therefore, in this paper, we use MD&A as textual information. Figure 1 shows the distribution of the number of Chinese characters in MD&A.

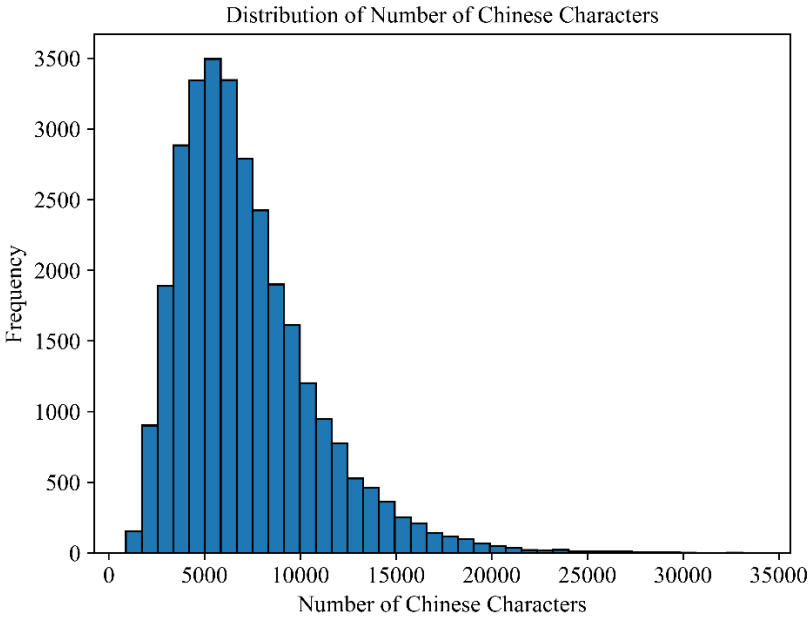

Figure 1. Distribution of Number of Chinese Characters.

As can be seen from Figure 1, most of the texts are concentrated in 2000 - 15000 words, a small number are distributed in 0-2000 and 15000-25000 words, and very few are distributed in more than 25000 words. This is too many words for BERT, the most used natural language processing model today. Since BERT can only support a maximum input of 512 characters, a new and effective text processing method is urgently needed.

2.2 Text processing

The processing methods of financial text can be roughly divided into two kinds, one is based on statistical methods, some scholars will construct a sentiment dictionary in the financial field, such as Shixuan Li [4] constructed a sentiment dictionary in the Chinese financial field, and then other scholars based on a particular dictionary for the statistics of different kinds of feature words such as negative, positive, neutral, etc., and represent the text according to the word frequency of the feature words [5]; another One is a deep learning method, which trains the text by neural network, embedding words into a low-dimensional space, and each word can be represented by a low-dimensional real-valued vector, and this word representation model is called word embedding [6]. Since the word embedding model can identify complex syntactic and semantic relationships, this paper uses the word vector model to process the text.

Specifically, we use the BERT model. The BERT model is a word vector model proposed by Google [7] to efficiently extract linguistic features from textual data. The architecture of BERT consists of multiple Encoder layers of Transformers [8].The input consists of two sentences.

Firstly, a special identifier will be added at the beginning of the first sentence [CLS], a [SEP] is added after the first sentence to indicate the end of the first sentence, and a [SEP] is added at the end to indicate the end of the second sentence. Since the number of characters in most MD&A texts exceeds the maximum number of characters allowed by BERT, we split the MD&A into sentences, enter the sentences into BERT separately, and represent them as a 3D one matrix as follows.

Assuming that the MD&A document consists of n sentences, the MD&A text can be represented as:

$$W = \{s_1, s_2, \ldots\ldots, s_n\}$$

where each sentence consists of a number of words, i.e. the sentence s_i can be represented as:

$$s_i = \{w_{i1}, w_{i2}, \ldots\ldots, w_{ik_i}\}$$

Each sentence s_i in the MD&A text is processed by adding [CLS] identifier at the beginning of the sentence and [SEP] identifier at the end of the sentence, and it is transformed into a format that can be processed by the BERT model and inputted into the BERT model, and a word vector is obtained for each word, and each word vector in the sentence s_i is connected to obtain the word embedding matrix sd_i of the sentence s_i, which is shaped as {k,d } Where d is the dimension of the word vector, the dimension of the word vector obtained by BERT is 768 dimensions, k is the number of words contained in sentence s_i.

Each sentence in the MD&A text W is represented as a word-vector matrix, then W can be represented as a set of multiple word-vector matrices $\{n, sd_i\}$, where n denotes as the number of sentences included in W.

3 Models

The core idea of Hierarchical Attention Networks (HAN) [9] is to use a hierarchical attention mechanism to improve the performance of text classification. At the Word Attention layer (Word Attention), the model usually uses a bidirectional GRU to convert the words into a vector representation of fixed dimensions, and then computes a weight vector through the attention mechanism, which will be used to weight and sum the word vectors to obtain a vector representation of the sentence. At the sentence level attention layer (Sentence Attention) the model evaluates the importance of each sentence and again through the attention mechanism computes a weight vector which is used to perform a weighted sum over the sentence vectors to obtain a vector representation of the document. Finally, the model inputs the vector representation of the document into a fully connected layer for classification. This approach solves the long text classification problem, but loses sequence information.

Recurrent Neural Network (RNN) is a class of neural networks specifically designed to process time-series data samples. But RNN also has some disadvantages. When the sequence is too long, it tends to cause the gradient to disappear, which makes the parameter update only capture local dependencies, and cannot capture long-term correlations or dependencies between sequences. In addition, as the distance increases, RNN cannot effectively utilise historical information.

Inspired by HAN and RNN, this study proposes a Recurrent Attention Network (RAN) that not only makes full use of the information of long texts, but also captures the sequence information of the text sufficiently to better understand the text and predict the future values of the sequence. The computational process of the RAN model is shown in Figure 2.

The word vector matrix sd_i of the ith sentence and the output obtained from the previous layer are jointly fed into the self-attention network, and an output is obtained after computation. This output and the i+1th word vector matrix sd_{i+1} are jointly used as input to the next self-attention network until all the word vector matrices are computed. The output of the last self-attention layer is computed by the fully connected neural network to get the final result.

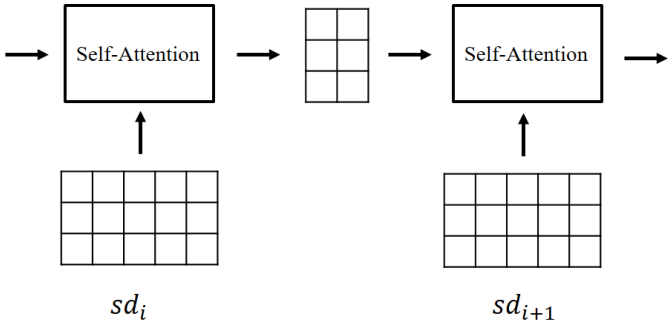

Figure 2. RAN model.

3.1 Self-Attention Network Computing

Suppose the input sentence is s $= \{w_1, w_2, \ldots\ldots, w_k\}$, the input word vector i_i corresponding to each word w_i is obtained after BERT processing, assuming that the dimension of i_i is d-dimensional, then the input of the self-attention network I $= \{i_1, i_2, \ldots\ldots, i_k\}$ is a $k * d$ dimensional matrix. Then according to the three parameter matrices W_Q、W_K、W_V by $Q = I \times W_Q$、$K = I \times W_K$、$V = I \times W_V$ the three matrices of Q, K, V are calculated as query vector matrix Quary, queried vector matrix Key, and value matrix Value respectively. Finally the output of self attention is obtained after self attention formula which is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right)V$$

where $Softmax$ is the normalised exponential function and $\sqrt{d_k}$ denotes the open square of the vector dimension of Quary and Key. The final $Attention$ value obtained is a matrix.

3.2 Recurrent Attention Network Computing

The result obtained from the computation of the self-attention network is matrix transformed so that it is spliced with the word vector matrix of the next sentence to obtain a new input, which is fed into the self-attention network again. The significance of using the same parameter matrix for the self-attention network is that the parameter space does not explode due to the increase in the length of the sequence.

4 Experiments

This section focuses on the experimental analysis of the proposed Recurrent Attention Network (RAN) model for financial distress prediction. In this section, we first present the dataset, baseline model, and evaluation metrics used during the experiments, and then assess the advantages of our model over various state-of-the-art models for financial distress prediction.

4.1 Dataset

This study uses the text of the firm's annual report in year T to predict whether the firm is in financial distress in year T+3. This is because in China, two consecutive consecutive fiscal years with negative net profits (losses) are labelled as STs, so predicting the financial position after three years is more meaningful.

Since the number of firms on ST each year is much smaller than the total number of firms, this study uses random undersampling from a sample of normal firms to create a 1:1 pairing with a sample of ST firms.

4.2 Baselines

In order to assess the validity of the RAN model, we compare our method with four other financial distress forecasting methods. Further details of these methods are described below:

(1) W2V_avg [2], the text is segmented and then trained by Word2Vec model to obtain word vectors, which are averaged in the text to obtain the text vectors as features for financial distress prediction.

(2) BERT_tfidf [3], the words in the text are passed through the BERT model to obtain word vectors and weighted by the tf-idf algorithm to obtain document vectors as features for financial distress prediction.

(3) BERT-LSTM [10], the word vectors obtained from BERT are fed into LSTM, and the output of the last LSTM neural cell is taken as a feature for financial distress prediction.

(4) HAN [9], a neural network for document classification, has as its central feature the application of two levels of attentional mechanisms at the word and sentence levels. At the word level, the model focuses on those words that are most important to the sentence representation; at the sentence level, the model focuses on those sentences that are most important to the overall document representation. This hierarchical attention mechanism allows the model to differentially engage with increasingly important content in the construction of the document representation.

4.3 Assessment of indicators

In this study, accuracy, AUC and KS were used as three evaluation metrics to assess the performance of the prediction model.

Accuracy (ACC) is the proportion of correctly predicted samples to the total number of samples.

AUC is the area under the subject operating characteristic curve (ROC, Receiver Operating Characteristic Curve) enclosed with the coordinate axis.

KS (Kolmogorov-Smirnov) is used to measure the ability of the classifier to distinguish between positive and negative samples. It calculates the maximum vertical distance between TPR and FPR on the ROC curve.

4.4 Experiments and analyses

In order to analyse the effectiveness of RAN models in financial distress prediction, we divided the dataset into a training set and a test set, where the training set is 80% of the dataset and the test set is 20% of the dataset. We examined the prediction performance of various models under different methods for 20 experimental results and took the average value as the experimental results, as shown in Table 2.

Table 2: Results of experiments in year T+3.

	RAN	W2V_avg	BERT_tfidf	BERT-LSTM	HAN
AUC	**0.8462**	0.7698	0.7888	0.8253	0.8392
ACC	**0.7608**	0.6850	0.7099	0.7452	0.7533
KS	**0.5588**	0.4168	0.4600	0.5401	0.5520

In order to show the metrics of different models more intuitively and to facilitate the comparison of the prediction performance of different models, we compared the results of 20 experiments using box-and-line plots, as shown in Figure 3.

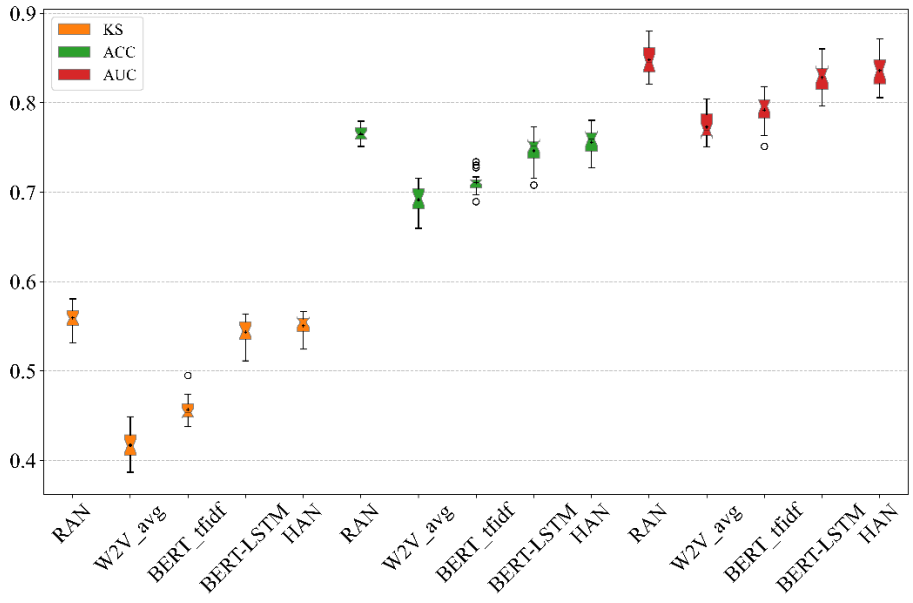

Figure 3. Box plot predicting financial distress in T+3 years.

From the results, it can be seen that the method of inputting the obtained word vectors into a neural network is more effective as compared to the simple method of word vector counting, due to the fact that this method captures the information in the document better. It can also be

found that using our proposed RAN approach achieves the best results in all the evaluation metrics and can be effective in predicting financial distress.

4.5 Robustness experiments

To verify the robustness of the above results, we conducted an additional set of experiments. The prediction objective of the experiments was changed to predicting whether a person is in financial distress in year T+5 based on the text of the annual report in year T, while maintaining a 1:1 sample ratio. We conducted 20 separate experiments for each group.

Table 3 shows the results of the experiments using year T data to predict whether or not to fall into a financial crisis in year T+5. Similar to previous findings, the RAN still exhibits the best text prediction performance, and therefore the RAN model is robust.

Table 3: Results of experiments in year T+5.

	RAN	W2V_avg	BERT_tfidf	BERT-LSTM	HAN
AUC	**0.755**	0.6751	0.6950	0.7208	0.7486
ACC	**0.7002**	0.6334	0.6289	0.6426	0.5923
KS	**0.4195**	0.2900	0.3140	0.3523	0.4113

5 Conclusions

In this paper, we try to extract features from the financial data of listed companies from the perspective of annual reports using recurrent attention networks, with a view to achieving the purpose of accurately predicting the company's financial distress. Through in-depth research and experimental validation, we constructed a financial distress prediction model based on recurrent attention networks and successfully applied it to actual data. The experimental results show that the model performs well in extracting key information from annual reports and effectively improves the accuracy and efficiency of financial distress prediction. This research not only provides new ideas and methods for financial distress prediction of listed companies, but also provides powerful support for risk management and investment decision-making in the financial market.

References

[1] P. Gandhi, T. Loughran, and B. McDonald, “Using annual report sentiment as a proxy for financial distress in US banks,” Journal of Behavioral Finance, vol. 20, no. 4, pp. 424–436, 2019.

[2] F. Mai, S. Tian, C. Lee, and L. Ma, “Deep learning models for bankruptcy prediction using textual disclosures,” European Journal of Operational Research, vol. 274, no. 2, pp. 743–758, Apr. 2019, doi: 10.1016/j.ejor.2018.10.024.

[3] C. Jiang, X. Lyu, Y. Yuan, Z. Wang, and Y. Ding, “Mining semantic features in current reports for financial distress prediction: Empirical evidence from unlisted public firms in China,” International Journal of Forecasting, vol. 38, no. 3, pp. 1086–1099, Jul. 2022, doi: 10.1016/j.ijforecast.2021.06.011.

[4] S. Li, W. Shi, J. Wang, and H. Zhou, “A Deep Learning-Based Approach to Constructing a Domain Sentiment Lexicon: a Case Study in Financial Distress Prediction,” Information Processing & Management, vol. 58, no. 5, p. 102673, Sep. 2021, doi: 10.1016/j.ipm.2021.102673.
[5] B.-H. Nguyen and V.-N. Huynh, “Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach,” Journal of the Operational Research Society, vol. 73, no. 1, pp. 102–121, 2022.
[6] M. Stevenson, C. Mues, and C. Bravo, “The value of text for small business default prediction: A deep learning approach,” European Journal of Operational Research, vol. 295, no. 2, pp. 758–771, 2021.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018..
[8] A. Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” presented at the Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.
[10] A. G. Katsafados, G. N. Leledakis, E. G. Pyrgiotakis, I. Androutsopoulos, and M. Fergadiotis, “Machine learning in bank merger prediction: A text-based approach,” European Journal of Operational Research, vol. 312, no. 2, pp. 783–797, Jan. 2024, doi: 10.1016/j.ejor.2023.07.039.