Applied Science and Biotechnology Journal for Advanced Research e-ISSN: 2583-553X
Peer Reviewed Journal
Volume-3, Issue-3 (May 2024), Page: 45-49 https://doi.org/10.5281/zenodo.12642108

Financial Fraud Detection in Listed Companies Using Deep Learning and Textual Emotion Analysis

Neha Romanenko[1], Kritika Sharma[2] and Siddharth Verma[3]
[1]Financial Risk, Indian School of Business (ISB), Hyderabad, India
[2]Business Administration,Indian Institute of Management (IIM) Bangalore, India
[3]Electronic Information Engineering, Indian Institute of Technology (IIT) Kanpur, India

[1]Corresponding Author: Gupta889@gmail.com

Received: 23-04-2024 Revised: 10-05-2024 Accepted: 29-05-2024

ABSTRACT
Financial fraud of listed companies refers to the bad faithless behaviour that improperly distorts accounting information, which hurts the company's management, economic development and social interests. At present, the existing research mainly focuses on financial digital data, while the exploration of text information and deep learning algorithms is relatively small. Therefore, this paper proposes a financial fraud identification method for listed companies based on deep learning and integrated text-emotional features. Firstly, the financial index is preprocessed, and then the Bi-LSTM model is used to extract the emotional features of the stock review text. Subsequently, a residual-cross-convolutional (RCC) parallel network structure is used to identify financial fraud. The network simultaneously uses a Residual network, Cross network, Convolutional network and long short-term memory network to extract the characteristics of financial fraud in a parallel way. It obtains the final recognition result through batch standardisation and a full connection layer.

Keywords: financial fraud identification, deep learning, text sentiment analysis, listed companies

I. INTRODUCTION

Financial fraud, or financial fraud, refers to the behavior of making the company's financial data distorted or false by improper means in order to seek improper benefits. This kind of behavior has seriously damaged the company's operation and social interests, including fabricating or concealing accounting information, falsely reporting balance sheets and profit statements, and adjusting account books without authorization. With the development of global economic integration and securities market, financial fraud of listed companies has become a global problem and shows a trend of continuous growth. Financial fraud not only destroys the trust of investors in enterprises, but also affects the interests of shareholders and the stability of the overall market, and has a serious impact on the financial revenue and social security of the society.

Although the country has introduced a series of strict regulations and measures to crack down on financial malpractices in listed companies, there are still many enterprises taking improper measures due to the pursuit of private interests. Most of the existing researches focus on traditional statistical methods and some machine learning algorithms, and rarely involve deep learning technology, and ignore the potential role of text information in the identification of financial fraud.

This paper aims to build a new financial fraud identification model based on deep learning by integrating the financial statement indicators of listed companies and the emotional characteristics of stock review texts. Specifically, firstly, the financial indicators and text are preprocessed and sentiment analyzed, which are spliced into the input of neural network; Secondly, three parallel network modules, including residual network, cross network and convolutional network, are used to extract financial fraud characteristics from different angles. Finally, by combining the output of the three networks through batch standardization and full connection layer, the final recognition result is obtained.

The purpose of this paper is to improve the ability of accurate identification of financial fraud in listed companies, and make contributions to early detection and prevention of financial fraud, so as to reduce its negative impact on the market and society.

II. RELATED WORK

2.1 Traditional Methods based on Statistics

Research on financial fraud identification of listed companies can be divided into three categories: traditional methods based on statistics, methods based on machine learning and methods based on neural networks. In one study, 50 enterprises involved in fraud and 50 enterprises not involved were selected as samples, 10 representative indicators were selected as the influence factors of fraud, and the financial fraud identification model was obtained by using Logistic regression. Yang Guijun et al. [4] proposed a financial fraud identification method combining Benford's Law and Logistic

model. After simulating and studying the financial data of Chinese listed companies, it was found that the Logistic model containing the Benford factor had a higher accuracy rate. Traditional methods based on statistics are often based on Logistic regression, and the identification performance is usually low. Moreover, most studies use accuracy as an evaluation index to measure the model performance. This can not properly evaluate problems such as the imbalance of positive and negative samples, such as financial fraud.

In most cases, Logistic regression and other methods are mainly used to select a small number of financial indicators to predict financial fraud. For example, Hui Wenjie used Logistic regression models to identify financial fraud. Still, these methods usually had low accuracy and failed to deal well with the problem of unbalanced positive and negative samples.

2.2 Machine Learning-Based Approach

On the financial fraud data set of European listed companies, we use a variety of integrated learning algorithms to detect financial fraud, including random forest, decision tree, CatBoost, and other algorithms, and we finally find that random forest performs best compared with other algorithms. Patel et al. [7] selected the data from the Bombay Stock Exchange, used data mining technology to select 10 important indicators, and selected 86 fraudulent companies and 92 non-fraudulent companies of manufacturing companies. Finally, the results of the random forest model were the best, and at the same time, the random forest model was improved to achieve better performance. Purda et al. [8] used text information of financial annual reports and interim reports of listed companies to distinguish fraud reports from real reports. The model used a text-based method and an SVM algorithm. Fraudulent reports can also be identified from a series of reports issued by a company. Dong et al. [9] also used the text content in financial statements. They developed a systematic text analysis framework for identifying financial fraud under the guidance of the theory of systemic functional linguistics. The framework extracts word-level and document-level features as input to Liblinear support vector machine. It achieves high recognition accuracy in financial statement data sets of listed companies in the United States. Compared with traditional methods, the method based on machine learning greatly improves recognition performance, and a few scholars have found the role of text information in identifying financial fraud [8-9]. Specific methods include SVM, decision tree, random forest and other algorithms to identify financial fraud. Compared with traditional statistical methods, these methods greatly improve and can effectively process a variety of financial data features. For example, Patel's research on the data of the Bombay Stock Exchange by using a random forest model has shown a good recognition effect.

2.3 Method based on Neural Network

In recent years, neural network-based methods have attracted much attention in the field of financial fraud identification, and various advanced neural network models such as multi-layer feedforward neural networks (MLFF), probabilistic neural networks (PNN), long short-term memory networks (LSTM), and Transformer have demonstrated superior performance in utilising text information and complex financial indicators.

1. Multi-Layer Feedforward Neural Network (MLFF) - as a classical deep learning model, it is widely used to deal with nonlinear relationships. In financial fraud identification, MLFF can effectively learn and model complex associations and trends in financial data. For example, it can analyze the complex interactions between revenue, cost, profit and other metrics in a company's financial statements to identify potential abnormal patterns and fraud.

2. Probabilistic Neural Networks (PNN) - which focus on modeling complex probability distributions, are unique in their ability to deal with uncertainty and risk in financial data. The application of PNN in financial fraud identification includes modeling and identifying probabilistic fraud patterns, and effectively capturing and predicting potential financial anomalies by analyzing the probabilistic relationship between different financial indicators.

3. Emerging Neural Network Models Such as Long Short-Term Memory Networks (LSTM) and Transformers perform well in processing natural language and text data. In financial fraud identification, these models can extract important information from financial statements, management discussion and analysis (MD&A) and other unstructured text data and effectively identify patterns and trends hidden behind large amounts of data. For example, LSTM can process time series data through the design of its memory unit, which is suitable for analyzing abnormal changes in the company's historical financial data. With its self-attention mechanism, Transformer can better understand and leverage relevant information in text data, making it adaptable to complex financial reporting and commentary text.

To sum up, the neural network-based approach not only shows the ability to deal with the challenges of financial fraud identification technically but also makes remarkable progress in application. With the continuous advancement of deep learning technology and the improvement of data processing capabilities, these methods are expected to provide more reliable and efficient solutions for financial fraud prevention and detection in the future.

III. FINANCIAL FRAUD IDENTIFICATION BASED ON DEEP LEARNING AND TEXT EMOTION MODEL

3.1 Model Structure

This paper aims to identify financial fraud in listed companies by using deepness. The model is trained and predicted. The expression ability of deep learning is stronger than that of machine learning and integrated learning, and it can better

learn the relationship between features, and has a better effect on the identification of financial fraud. The model structure of this paper is shown in Figure 1.

The RCC parallel network model structure proposed in this paper can be divided into the following levels:

(1) Input layer: input financial statement data and stock evaluation data;

(2) Pre-processing layer: Pre-processing financial statement data, emotion extraction of stock evaluation data, oversampling and standardisation after splicing;

(3) Feature extraction layer: Use parallel residual network module, cross network module and convolutional network module to extract financial fraud feature information of financial statements and text emotion;

(4) Full connection layer: The output of multiple networks in the parallel network layer is spliced, and the relationship between multiple networks and the target result is deepened through batch standardization and full connection;

(5) Output layer: The output of the model is obtained through Sigmoid binary classification

In this article, the research team explored the predictive analysis of customers' online purchasing behavior from the data of the customer purchase prediction contest under the consumer finance scenario held by the Credit Card Center of China Merchants Bank. The following are the main methodological parts:

3.2 Pretreatment Layer

Financial statements contain more financial indicators. The study found that some of the money Business indicators are more important than other financial indicators, and some indicators negatively affect classification accuracy. Therefore, the choice of financial indicators is crucial for any learning algorithm. Because there are many missing values in the financial statement data, it is necessary to clean the data to obtain higher quality data. The feature values with missing values greater than 30% are deleted. Then, this paper uses embedding method and ensemble learning model for feature selection, while taking into account existing research, selects 50 financial indicators in financial statements as the features of the financial fraud identification model, and the selected features can measure the financial status of listed companies from multiple perspectives. The data includes financial statements of multiple years and is grouped into listed companies. For the missing value of the class type value, the intra-group mode will be used to fill in; For the missing values of continuous values, the method of step interpolation will be used to fill in the group. The data is then sorted by year and time, and the final result contains no missing data

3.3 Parallel Network Layer

The parallel network layer inputs the input vectors to the residual network module,In the network module and convolutional network module, three network modules are used to extract financial fraud characteristics simultaneously, fully using the advantages of different neural networks. In the study of financial fraud, some financial indicators may have little or no correlation with financial fraud, and the deep learning model has high complexity and strong fitting ability, so it is easy to overfit financial fraud data sets. At the same time, if the number of layers in the network is too deep, the problem of gradient disappearance or gradient explosion will easily appear. Therefore, to optimize the model's performance again based on feature engineering, this paper adds the residual structure [17] based on a deep neural network to form a residual block, and multiple residual blocks are connected together to form the residual network module of this model.

IV. CONCLUSION

In conclusion, this paper proposes a novel approach to identifying financial fraud in listed companies by integrating deep learning techniques with text-emotional features extracted from stock review texts. The research addresses the limitations of existing methods that predominantly focus on traditional statistical and machine learning approaches while neglecting the potential of deep learning and textual information in fraud detection.

The proposed method demonstrates robust capabilities in identifying financial irregularities by leveraging the Bi-LSTM model for sentiment analysis and a residual-cross-convolutional (RCC) parallel network structure. This innovative framework utilizes a combination of residual, cross, convolutional, and long short-term memory networks to effectively extract and analyze fraudulent patterns from financial data and sentiment-laden textual information.

The findings highlight the significant potential of deep learning methodologies in enhancing the accuracy and efficiency of financial fraud detection systems. By integrating financial indicators with sentiment analysis from textual sources, the proposed model improves detection capabilities and contributes to early prevention and mitigation of financial fraud risks in listed companies.

REFERENCES

1. Yuan, J., Lin, Y., Shi, Y., Yang, T., & Li, A. (2024). Applications of artificial intelligence generative adversarial techniques in the financial sector. *Academic Journal of Sociology and Management*, *2*(3), 59-66.
2. Lin, Y., Li, A., Li, H., Shi, Y., & Zhan, X. (2024). GPU-Optimized image processing and generation based on deep learning and computer vision. *Journal of Artificial Intelligence General science (JAIGS)*, *5*(1), 39-49.

3. Chen, Zhou, et al. (2024). Application of cloud-driven intelligent medical imaging analysis in disease detection. *Journal of Theory and Practice of Engineering Science 4*(05), 64-71.
4. Wang, B., Lei, H., Shui, Z., Chen, Z., & Yang, P. (2024). *Current state of autonomous driving applications based on distributed perception and decision-making*.
5. Zhan, X., Shi, C., Li, L., Xu, K., & Zheng, H. (2024). Aspect category sentiment analysis based on multiple attention mechanisms and pre-trained models. *Applied and Computational Engineering, 71*, 21-26.
6. Wu, B., Xu, J., Zhang, Y., Liu, B., Gong, Y., & Huang, J. (2024). Integration of computer networks and artificial neural networks for an AI-based network operator. *Applied and Computational Engineering*, *64*, 115-120.
7. Liang, P., Song, B., Zhan, X., Chen, Z., & Yuan, J. (2024). Automating the training and deployment of models in MLOps by integrating systems with machine learning. *Appl. Comput. Eng.*, *67*, 1–7. Available at: https://doi.org/10.54254/2755-2721/67/20240690.
8. Li, A., Yang, T., Zhan, X., Shi, Y., & Li, H. (2024). Utilizing data science and AI for customer churn prediction in marketing. *Journal of Theory and Practice of Engineering Science*, *4*(05), 72-79.
9. Wu, B., Gong, Y., Zheng, H., Zhang, Y., Huang, J., & Xu, J. (2024). Enterprise cloud resource optimization and management based on cloud operations. *Applied and Computational Engineering*, *67*, 8-14.
10. Xu, J., Wu, B., Huang, J., Gong, Y., Zhang, Y., & Liu, B. (2024). Practical applications of advanced cloud services and generative AI systems in medical image analysis. *Appl. Comput. Eng.*, *64*, 83–88. https://doi.org/10.54254/2755-2721/64/20241361.
11. Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024). Application of machine learning optimization in cloud computing resource scheduling and management. *Appl. Comput. Eng.*, *64*, 17–22. https://doi.org/10.54254/2755-2721/64/20241359.
12. Huang, J., Zhang, Y., Xu, J., Wu, B., Liu, B., & Gong, Y. (2024). Implementation of seamless assistance with Google Assistant leveraging cloud computing. *Appl. Comput. Eng.*, *64*, 170–176. https://doi.org/10.54254/2755-2721/64/20241383.
13. Shi, Y., Li, L., Li, H., Li, A., & Lin, Y. (2024). Aspect-Level sentiment analysis of customer reviews based on neural multi-task learning. *Journal of Theory and Practice of Engineering Science*, *4*(04), 1-8.
14. Shi, Y., Yuan, J., Yang, P., Wang, Y., & Chen, Z. (2024). Implementing intelligent predictive models for patient disease risk in cloud data warehousing. *Appl. Comput. Eng.*, *67*, 34–40. https://doi.org/10.54254/2755-2721/67/2024ma0059.
15. Zhan, T., Shi, C., Shi, Y., Li, H., & Lin, Y. (2024). Optimization techniques for sentiment analysis based on LLM (GPT-3). *Appl. Comput. Eng.*, *67*, 41–47. https://doi.org/10.54254/2755-2721/67/2024ma0060.
16. Li, Huixiang, et al. (2024). AI face recognition and processing technology based on GPU computing. *Journal of Theory and Practice of Engineering Science, 4*(05), 9-16.
17. Jiang, W., Qian, K., Fan, C., Ding, W., & Li, Z. (2024). Applications of generative AI-based financial robot advisors as investment consultants. *Appl. Comput. Eng.*, *67*, 28–33. https://doi.org/10.54254/2755-2721/67/2024ma0057.
18. Ding, W., Zhou, H., Tan, H., Li, Z., & Fan, C. (2024). *Automated compatibility testing method for distributed software systems in cloud computing*.
19. Haowei, Ma, et al. (2023). CRISPR/Cas-based nanobiosensors: A reinforced approach for specific and sensitive recognition of mycotoxins. *Food Bioscience*, *56*, 103110.
20. Fan, C., Li, Z., Ding, W., Zhou, H., & Qian, K. (2024). Integrating artificial intelligence with SLAM technology for robotic navigation and localization in unknown environments. *Appl. Comput. Eng.*, *67*, 22–27. https://doi.org/10.54254/2755-2721/67/2024ma0056.
21. Guo, L., Li, Z., Qian, K., Ding, W., & Chen, Z. (2024). Bank credit risk early warning model based on machine learning decision trees. *Journal of Economic Theory and Business Management, 1*(3), 24-30.
22. Li, Zihan, et al. (2024). *Robot navigation and map construction based on SLAM technology*.
23. Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing flight object trajectory and safety prediction based on SLAM technology. *Journal of Theory and Practice of Engineering Science, 4*(05), 1-8.
24. Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. (2024). Immediate traffic flow monitoring and management based on multimodal data in cloud computing. *Appl. Comput. Eng.*, *71*, 1–6. https://doi.org/10.54254/2755-2721/71/2024ma0052.
25. Bi, Shuochen, Wenqing Bao, Jue Xiao, Jiangshan Wang, & Tingting Deng. (2024). *Application and practice of AI technology in quantitative investment*. arXiv preprint arXiv:2404.18184(2024).
26. Qian, K., Fan, C., Li, Z., Zhou, H., & Ding, W. (2024). Implementation of artificial intelligence in investment decision-making in the chinese a-share market. *Journal of Economic Theory and Business Management, 1*(2), 36-42.
27. Liang, P., Song, B., Zhan, X., Chen, Z., & Yuan, J. (2024). Automating the training and deployment of models in MLOps by integrating systems with machine learning. *Appl. Comput. Eng.*, *67*, 1–7. https://doi.org/10.54254/2755-2721/67/20240690.

28. Haowei, M. A., et al. (2023). Employing Sisko non-Newtonian model to investigate the thermal behavior of blood flow in a stenosis artery: Effects of heat flux, different severities of stenosis, and different radii of the artery. *Alexandria Engineering Journal*, *68*, 291-300.
29. Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. (2024). RNA secondary structure prediction using transformer-based deep learning models. *Appl. Comput. Eng.*, *64*, 95–101. https://doi.org/10.54254/2755-2721/64/20241362.
30. Xiao, J., Wang, J., Bao, W., Deng, T., & Bi, S. *Application progress of natural language processing technology in financial research*.
31. Liu, B., Cai, G., Ling, Z., Qian, J., & Zhang, Q. *Precise positioning and prediction system for autonomous driving based on generative artificial intelligence*.（10个）
32. Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. *Precision gene editing using deep learning: A case study of the CRISPR-Cas9 Editor*.
33. Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. *Applied and Computational Engineering*, *64*, 95-100.
34. He, Zheng, et al. *Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering*.
35. Song, J., Cheng, Q., Bai, X., Jiang, W., & Su, G. (2024). LSTM-based deep learning model for financial market stock price prediction. *Journal of Economic Theory and Business Management*, *1*(2), 43-50.
36. Ni, C., Zhang, C., Lu, W., Wang, H., & Wu, J. (2024). *Enabling intelligent decision making and optimization in enterprises through data pipelines*.
37. Lu, W., Ni, C., Wang, H., Wu, J., & Zhang, C. (2024). *Machine learning-based automatic fault diagnosis method for operating systems*.