2024 4th International Symposium on Computer Technology and Information Science (ISCTIS) | 979-8-3503-5456-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCTIS63324.2024.10699096

A FinBERT Framework for Sentiment Analysis of Chinese Financial News

Yuesheng Huang[1], Renyi Lin[1], Jiawen Li[1,2,*], Yiqi Li[3], Jiajia Chen[1], Jiaqi Liang[1], Guanyuan Feng[1], Jujian Lv[1]
[1]School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China
[2]Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, China
[3]School of Finance and Economics, Guangdong Polytechnic Normal University, Guangzhou, China
*Corresponding email: lijiawen@gpnu.edu.cn

***Abstract*-In the era of big data, sentiment analysis of the Chinese stock market has become a research hotspot. As a vital reflection of market sentiment, financial news has a significant impact on stock market trends. In this paper, we propose a framework based on the FinBERT pre-trained model optimized for financial news sentiment analysis in China. In particular, the model utilizes the Bidirectional Encoder Representations from Transformers (BERT) to deal with semantics, syntactic features, and specialized vocabulary of financial texts through self-supervised learning. The fine-tuning stage further optimizes the model, so that it can accurately predict sentiment tendencies. Experimental results show that the proposed model outperforms others on the sentiment classification task, with an accuracy of up to 94.52%. Consequently, the model can provide participants in the financial industry with a valuable reference of market trends, which helps investors and analysts make decisions properly.**

Keywords-FinBERT; Financial News; Sentiment Analysis

I. Introduction

In the expansive domain of global financial markets, the Chinese stock market has garnered considerable attention due to its distinctive characteristics. Financial markets are subject to various influencing factors, among which the emotional sentiments of market participants play a crucial role. Financial news, as a vital conduit of information, contains emotional responses and public perspectives concerning pertinent events, exerting a profound influence on the trajectory of financial markets [1]. With the advent of the big data era, the fields of natural language processing and machine learning have become increasingly integral to sentiment analysis within China's financial sector, thereby augmenting the sophistication of text sentiment analysis techniques. Nonetheless, the scarcity of annotated data in this context poses a constraint on the enhancement of sentiment classification accuracy for domestic stocks [2]. To this end, researchers have embarked on extensive and profound investigations within this domain. For example, Xiao and Ihnaini [3] introduced a pre-trained FinBERT model tailored for the analysis of financial texts, which demonstrated enhanced generalization capabilities within the financial sector. Daudert [4] achieved good performance by refining the BERT model which helped the evolution of sentiment analysis. Liapis and Kotsiantis [5] delved into the application of the FinBERT model in the context of multi-label sentiment analysis for financial texts, effectively broadening the scope of sentiment analysis applications. Finally, Kim et al. [6] developed a hybrid model that includes FinBERT and Long Short-Term Memory Network (LSTM) for sentiment analysis in finance, proposing several innovative enhancements to the model. As seen, these contributions have collectively propelled the field forward, offering nuanced insights into the complexities of sentiment analysis within the financial industry.

Now, to make the sentiment analysis of financial news more efficient, following the needs of financial news texts, this paper introduces the FinBERT pre-trained language model, which is a customized model based on BERT specifically tailored to the conditions of sentiment analysis in the financial domain, with a deeper understanding of the industry's discourse and sentiment expression. As for the method validation, we apply the FinBERT model to the sentiment analysis of Chinese financial news, aiming to promote the development of financial news information processing and provide participants in the financial industry with more valuable references for tendency.

This paper is organized as follows: In Section II, we present the dataset evaluated and the methodology employed. Section III provides a detailed account of the experimental findings. Lastly, Section IV summarizes the key conclusions drawn from our analysis.

II. Proposed Method

A. BERT

BERT is a pre-trained model based on Transformer, which has achieved remarkable results in the field of Natural Language Processing (NLP). It mainly uses the encoder part of the transformer as its infrastructure, the specific encoder framework of which is illustrated in Figure 1. The transformer encoder consists of multiple identical layers stacked on top of each other, each of which contains a self-attention mechanism and a feed-forward neural network. Unlike traditional RNN or LSTM-based models, the transformer can process all positions of the input sequence in parallel, which increases the training speed according.

*This work was supported in part by the Guangzhou Science and Technology Plan Project under Grants 2024B03J1361, 2023B03J1327, and 2023A04J0361 in part by the Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security under Grant MIMS22-02, in part by the Guangdong Province Ordinary Colleges and Universities Young Innovative Talents Project under Grant 2023KQNCX036, in part by the Special Fund for Science and Technology Innovation Strategy of Guangdong Province (Climbing Plan) under Grant pdjh2024a226, and in part by the Research Fund of Guangdong Polytechnic Normal University under Grants 22GPNUZDJS17 and 2022SDKYA015.

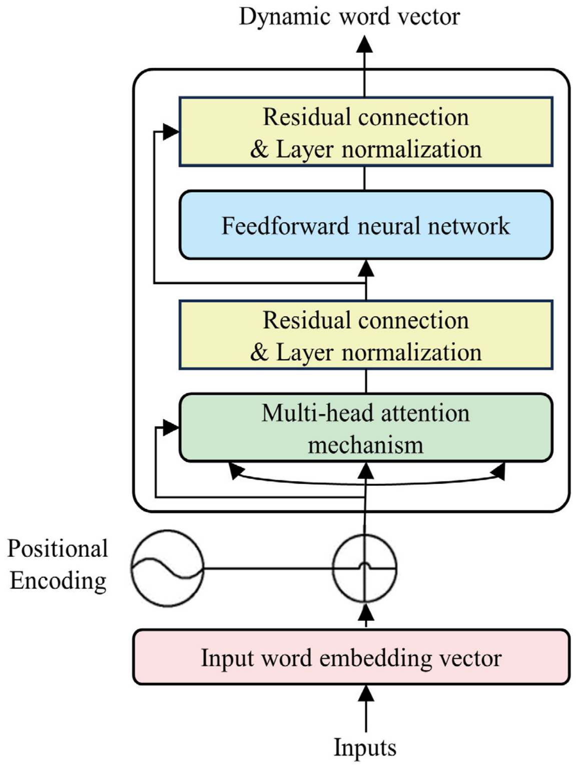

Figure 1. The architecture of the transformer encoder in the BERT

1) Masked Language Modeling (MLM)

During the training process, BERT randomly masks some of the tokens in the input sequence and tries to predict these masked tokens. This requires the model to have a deep understanding of the context because the prediction of the masked tokens needs to depend on their surrounding tokens.

2) Next Sentence Prediction (NSP)

To understand the relationship between sentences, BERT accepts pairs of sentences as input and predicts whether the second sentence is a successor to the first, which helps the model to understand sentence-level relationships. After pre-training is complete, BERT can be used for a variety of NLP tasks. By simply adding task-specific output layers (e.g., a classification layer or a sequence generation layer) and fine-tuning them on task-specific data, BERT can achieve excellent performance on a variety of NLP tasks. Since BERT is pre-trained on a large amount of unlabeled text, it can capture universal laws and patterns of language. This enables BERT to achieve excellent results on a variety of NLP tasks, especially those lacking a large amount of labeled data. The bi-directionality of BERT allows it to capture richer contextual information that improves the performance of the mode, so the transformer structure of BERT assists it in processing all positions of the input sequence in parallel, improving the training speed.

B. FinBERT-Sentiment

In this paper, we use a FinBERT-based variant model, FinBERT-Sentiment, which is specially optimized and fine-tuned for the task of sentiment analysis of Chinese financial news texts. Its architecture, principles, and advantages for the Chinese financial news text sentiment analysis are described in detail below.

The architecture of FinBERT-Sentiment is derived from the BERT model, which realizes the accurate judgment of the sentiment tendency of Chinese financial news texts by introducing expertise in the financial domain and fine-tuning of specific tasks. Its core architecture is displayed in Figure 2.

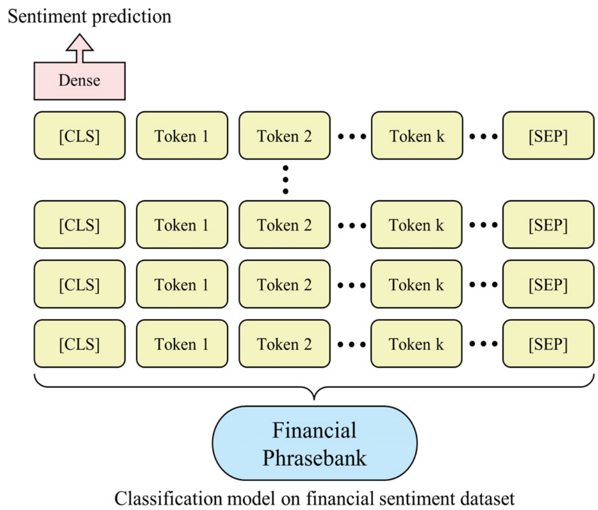

Figure 2. The proposed FinBERT-Sentiment model

As seen, the main role of the input layer is to receive Chinese financial news text as input and convert the text into numerical vectors that can be processed by the model through steps such as word splitting and word embedding. Here, the BERT encoder adopts a multi-layer transformer encoder structure, which beneficially captures the contextual information in the text through the self-attention mechanism, and generates a vector representation of the text accordingly. Then, a sentiment classification layer is added on top of the BERT encoder, which helpfully takes the output of the BERT encoder as an input and performs sentiment classification through a fully connected network. The sentiment classification layer contains multiple neurons, each corresponding to a sentiment category. The training process of the FinBERT-Sentiment model is mainly divided into two phases: pre-training and fine-tuning:

1) Pre-training

In the pre-training phase, we employ unlabeled Chinese financial news text data to train the model. Utilizing self-supervised learning, the model can learn the semantic and syntactic features of financial texts, as well as specialized vocabulary and expressions in the financial domain. The training in this phase gives the model the basic ability to process financial texts.

2) Fine-tuning

In the fine-tuning stage, we apply a Chinese financial news text dataset with sentiment labels to train the model. Through optimization methods such as back-propagation algorithm and gradient descent, the model continuously adjusts the weights of the sentiment classification layer, which promotes the model to more accurately predict the sentiment tendency of financial texts. Thus, after fine-tuning, the FinBERT-Sentiment model achieves superior performance on the task of analyzing the sentiment of Chinese financial news texts.

Besides, the FinBERT-Sentiment has several advantages in Chinese financial news text sentiment analysis. First, since FinBERT-Sentiment has learned the features of financial texts in the pre-training phase and optimized for the sentiment analysis task in the fine-tuning phase, the model has high accuracy. Second, the FinBERT-Sentiment is based on the BERT architecture, which exhibits better generalization ability. Even when facing unseen financial text data, the model can perform robust sentiment analysis with its powerful representation learning capability. Finally, the FinBERT-Sentiment is not only suitable for Chinese financial news text sentiment analysis but also can be extended to other financial text-related NLP tasks, such as company report analysis, and public opinion monitoring.

Furthermore, we take a deeper look at the problem of sentiment analysis and redefine it from the traditional view of an answer generation problem to a multi-category classification. By adopting the framework of multi-category classification, we can accurately identify the sentiment tendencies embedded in a text, be they positive, negative, or neutral. This approach has been widely explored and validated in numerous NLP studies, demonstrating its robustness and precision in sentiment analysis. Compared to traditional answer generation problems, which often simplify sentiment into binary categories, multi-category classification is more capable of capturing the subtleties and complexities of textual sentiment. This nuanced understanding allows for a more granular analysis, distinguishing between varying degrees of sentiment within the same category. Hence, this method significantly improves the accuracy and reliability of sentiment analysis, leading to more insightful and actionable results in applications such as customer feedback analysis, social media monitoring, and opinion mining.

In (1), we elaborate on the organization of the training sequence X that is input to the transformer model. To fully utilize the powerful performance of the model, we will ensure that the training sequences are diverse and representative while following certain rules and formats so that the model can effectively learn and recognize features from different sentiment categories. Such an organization will help to improve the generalization ability of the model and enable it to perform well in other real-world applications.

$$X = |BOS|\vec{q}_1|EOS| \tag{1}$$

where BOS denotes the sequence start marker and EOS is the sequence end marker.

C. Evaluation Metrics

In the classification task, accuracy is an important measure of the model's prediction performance, which indicates the percentage of correct predictions (either positive or negative) out of the total predictions. The formula for calculating the precision is (2):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

where TP means that the predicted value is 1 and the true value is also 1; TN means that the predicted value is 0 and the true value is also 0; FP means that the predicted value is 1 but the true value is 0; and FN means that the predicted value is 0 but the true value is 1.

Precision measures the proportion of samples predicted by the model to be positive cases that are actually positive cases (3):

$$Precision = \frac{TP}{TP + \mathrm{FP}} \tag{3}$$

The recall rate, also known as the true positive rate, indicates the proportion of all samples that are true positive examples that are predicted to be positive by the model (4):

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

III. Experiments

In this paper, the open-source financial news sentiment analysis dataset on https://gitcode.com/wwwxmu/Dataset-of-financial-news-sentiment-classification was evaluated based on positive and negative news headlines published by Vantage Information on Snowball. In detail, 7,046 news headlines were collected, including 5,147 positive news and 1,899 negative news, as an initial seed dataset for our study.

However, due to the relatively small amount of data in the initial dataset, a data expansion strategy was implemented, which relies heavily on the search and filtering functions of the search engine. Through the well-designed search criteria and screening standards, we effectively expanded the dataset, enhancing the reliability and representativeness of this study accordingly. The expanded dataset contains a total of 17,149 news data, each containing the following six key fields: date, company, ticker symbol, positive/negative tags, title, and body text. Among them, the total number of positive news items is 12,514, and the total number of negative news items is 4,635. This expanded dataset not only provides numerous samples for subsequent model training but also provides a solid foundation for in-depth research in the field of financial news sentiment analysis. It is particularly suitable for deep learning models, such as the BERT and LSTM, where they can make full use of the textual information and sentiment labels in the dataset for efficient feature extraction and pattern recognition, thus realizing accurate judgment of the sentiment tendency.

The source language is Chinese, containing the news date, company name, and ticker symbol. For the convenience of readers, we have translated the headline into English, but the model training uses the original data. The samples are listed in Table I.

TABLE I. Samples Of Seed Dataset In This Work

Date	Company (Translated from Chinese)	Ticker symbol	Positive/Negative	News Headline (Translated from Chinese)
2019/2/14	Yongxing Special Steel	002756	Positive	Net profit increased by 14.45% in 2018, proposed 10 shares for 10 dividends
2019/2/14	Dio Home Furnishings	002798	Positive	Net profit of RMB 382 million in 2018, representing a 6-fold increase
2019/2/14	Aibisen	300389	Positive	Net profit increased by 132% in 2018
2018/5/10	Xin Chao Energy	600777	Negative	Terminated restructuring, resumed trading on May 10th
2018/5/10	Anuoqi	300067	Negative	Terminated acquisition of Yancheng Dongwu equity, resumed trading on May 10th
2018/5/10	Teli A	000025	Negative	The second largest shareholder plans to reduce its shareholding by no more than 6%

TABLE II. Benchmarking Results Using Different Models

Model	Accuracy (%)	Precision (%)	Recall (%)
BERT [7]	90.59	91.48	88.69
CNN [8]	81.47	83.08	79.32
SVM [9]	72.91	74.82	70.97
This Work	94.52	95.26	93.41

Moreover, Table II shows the benchmarking results using different models. In the financial news sentiment classification task, the FinBERT model significantly outperforms others, including the original BERT, CNN, and Support Vector Machine (SVM), with an accuracy of 94.52%, a precision of 95.26%, and a recall of 93.41%, respectively. Such results suggest that the proposed FinBERT-Sentiment model owns a significant advantage in capturing the sentiment tendency of financial texts. The reason may be due to the domain-specific pre-training giving FinBERT-Sentiment an in-depth understanding of the linguistic characteristics and sentiment tendencies of the financial domain. Meanwhile, as a BERT-based deep learning model, the deep bi-directional representation capability enables it to effectively capture subtle sentiments and long-distance dependencies in the text, so the model parameters optimized for the task can further enhance the accuracy of the prediction. In this regard, if the pre-training data is highly correlated with the test set in terms of topic and style, it also enhances the model's adaptability and performance. Therefore, these factors help FinBERT to outperform others in the financial text sentiment classification task.

IV. Conclusion

In this paper, we propose a FinBERT model-based framework for analyzing the sentiment tendency of Chinese financial news. With a well-designed data extension strategy and optimized model parameters, the FinBERT-Sentiment model performs well in the sentiment classification task and significantly outperforms other models such as BERT, CNN, and SVM. The results demonstrate that the proposed method has a significant advantage in capturing the sentiment nuances of financial texts. In the future, we will further explore and integrate other emotion recognition techniques [10] to improve the model's accuracy and generalization capabilities, providing deep insights to participants in the financial domain.

References

[1] N. Jing, Z. Wu, and H. Wang, "A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction", Expert Systems with Applications, vol. 178, p. 115019, 2021.

[2] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text", Contemporary Accounting Research, vol. 40, no. 2, pp. 806–841, 2023.

[3] Q. Xiao and B. Ihnaini, "Stock trend prediction using sentiment analysis", PeerJ Computer Science, vol. 9, p. e1293, 2023.

[4] T. Daudert, "Exploiting textual and relationship information for fine-grained financial sentiment analysis", Knowledge-based Systems, vol. 230, p. 107389, 2021.

[5] C. M. Liapis and S. Kotsiantis, "Temporal convolutional networks and BERT-based multi-label emotion analysis for financial forecasting", Information, vol. 14, no. 11, p. 596, 2023.

[6] J. Kim, H. S. Kim, and S. Y. Choi, "Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: a FinBERT transformer model and LSTM", Axioms, vol. 12, no. 9, p. 835, 2023.

[7] M. Li, L. Chen, J. Zhao, et al. "Sentiment analysis of Chinese stock reviews based on BERT model", Applied Intelligence, vol. 51, pp. 5016–5024, 2021.

[8] N. Das, B. Sadhukhan, S. S. Bhakta, et al. "Integrating EEMD and ensemble CNN with X (Twitter) sentiment for enhanced stock price predictions", Social Network Analysis and Mining, vol. 14, no. 1, p. 29, 2024.

[9] S. A. Alanazi, A. Khaliq, F. Ahmad, et al. "Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques", International Journal of Environmental Research and Public Health, vol. 19, no. 15, p. 9695, 2022.

[10] J. W. Li, S. Barma, P. U. Mak, et al. "Single-channel selection for EEG-based emotion recognition using brain rhythm sequencing", IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2493–2503, 2022.