Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Enhancing Stock Market Prediction with Sentiment Analysis Using a BERT-based Model

Yifan Lin [1], Sen Wang [2, *]

[1] School of Life Sciences, Fudan University, Shanghai, China

[2] Overseas Chinese College, Capital University of Economics and Business, Beijing, China

* Corresponding author: 32021140016@cueb.edu.cn

Abstract. This study explores the prediction of stock price trends in the financial market, emphasizing the impact of investor sentiment and macroeconomic policies. Traditional research often uses mathematical, statistical, or deep learning methods to predict stock prices but overlooks the emotional factors in vast unstructured text data, such as financial news. This paper proposes a Bidirectional Encoder Representations from Transformers (BERT)-Transformer model that integrates sentiment analysis to enhance stock market prediction. Using financial news text data from the Oriental Fortune Network, the BERT model performs sentiment analysis to extract emotional polarity features. These features, combined with stock transaction data, are then input into a Transformer model to predict the index trends. The experimental results demonstrate a 60% accuracy in predicting stock indices' rise and fall, indicating the model's effectiveness while highlighting areas for improvement. The methodology details dataset preprocessing, the BERT-Back Propagation (BP) model structure, and how sentiment classification is combined with stock trading data for predictions. Performance comparisons with other prediction models confirm the benefits of incorporating sentiment analysis. The BERT-Transformer model's long-term memory capabilities make it a promising tool for financial time series predictions, offering new research ideas and methods for the financial field. Future research will aim to automate the sentiment feature labeling process to save time and improve efficiency.

Keywords: Sentiment Analysis; Stock Market Prediction; BERT-Transformer Model; Financial Market.

1. Introduction

In the financial field, the price trend of stocks receives widespread attention because it can affect the mood of investors and the macroeconomic policies of a country. Current research mainly uses mathematical, statistical or deep learning methods to predict future price trend of stocks based on structured data like stock market data. In the era of big data, a large amount of unstructured text data such as financial text data can be found on the internet. These data can create emotional factors that influence the rise and fall of the stock market [1]. Sentiment analysis on unstructured data is of profound value in predicting the price trend of stocks.

Sentiment analysis based on deep learning can extract emotional polarity features effectively. Zhang et al. pointed out that deep learning can learn deep representations or features in data and achieved optimal results compared with other machine learning methods [2]. This document summarized the deep learning technologies used in sentiment analysis, such as Word Embedding, Autoencoder, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and memory network and introduced some related methods and literature on sentence-level sentiment classification, such as Recursive Autoencoders Network, Dynamic CNN, CNN-multichannel, bidirectional LSTM, CNN-LSTM et al. [3-7]. Mikolov et al. proposed the C&W model and Skip-Gram model and used them for training to generate low-dimensional word vectors containing grammatical information and semantic information [8]. Inspired by the C&W model, Tang et al. proposed three neural networks used to learn emotional information word vectors and achieved results equivalent to the best existing methods based on manual features on the SemEval 2013

emotional data [9]. Le et al. constructed paragraph vectors under the framework of word2vec and applied paragraph vectors to sentiment classification [10]. These methods were different but shared the same general framework and there was not enough evidence to show which one is better. In 2018, Google released the milestone model Bidirectional Encoder Representations from Transformers (BERT). The results achieved by BERT refreshed 11 Natural Language Processing (NLP) tasks including sentiment analysis. It changed the research direction of natural language processing. BERT has two stages, which are pre-training and fine-tuning. The complexity of the pre-training model surpasses that of the original RNN and other algorithms. It has also been trained with billions of data [11]. Therefore, BERT can better fit the deep features of text and thus achieved optimal results in sentiment classification.

In order to better explore the stock market prediction value contained in financial news texts, this study took the stock data of 300 Shanghai and Shenzhen stocks from January 2015 to 2024 as the research object, and constructed an index prediction model BERT- Back Propagation (BERT-BP) based on financial text sentiment analysis. This model obtained financial news text data from Oriental Fortune Network, used the BERT model to perform sentiment analysis, extracted sentiment polarity features and then input the sentiment features and stock transaction data into the Transformer model to predict the price trend of 300 Shanghai and Shenzhen stocks. Model in research has achieved a 60% accuracy rate in predicting the rise and fall of stock indices, indicating that the model has a certain degree of effectiveness. However, there is still room for improvement, and we will continue to optimize the model to enhance its predictive accuracy. This study aims to provide a more comprehensive understanding of the factors affecting the price trend of stocks and to improve the accuracy of forecast models.

2. Methodology

2.1. Dataset Description and Preprocessing

The dataset used in this study included stock transaction data and news text data. For stock transaction data, this study chose the Crime Scene Investigation (CSI) 300 Index as the research object [12]. The technical analysis data used in this study included the day's trading volume, highest price, lowest price, percent change, opening price and closing price. These data only need to be normalized. The formula used for normalization is as follows:

$$x = \frac{x' - \min(x)}{\max(x) - \min(x)} \tag{1}$$

x' is the result of normalization, x is raw data, max(x) is the largest value among similar data and min(x) is the smallest valve among similar data. Table 1 and Table 2 demonstrated the original data and the normalized data respectively.

Table 1. Raw data

Day	Closing value	Opening value	Highest price	Lowest price	Trading value	Percent change
20231214	3351.96	3387.38	3399.14	3350.76	83.66k	-0.52%
20231213	3369.60	3418.64	3418.64	3369.60	95.29k	-1.67%
20231212	3426.80	3410.00	3427.55	3406.10	105.59k	0.21%

Table 2. Normalized data

Day	Closing value	Opening value	Highest price	Lowest price	Trading value	Percent change
20231214	0.16865	0.17339	0.15355	0.18095	0.06159	0.52683
20231213	0.17462	0.18359	0.16007	0.18743	0.07970	0.45157
20231212	0.19399	0.18077	0.16304	0.19986	0.09573	0.57460

The financial news text data used in this study came from the choice financial terminal provided by Oriental Fortune. The financial news data used in this study included financial news text data from the 'Shanghai Securities News' about the top 40 weighted stocks among the 300 Shanghai and Shenzhen stocks, from January 2, 2017 to December 30, 2023. There were 8837 items in total. Daily news was marked with bullish, bearish and no impact, which were 1, 2 and 0 respectively, according to whether the content conveyed by the daily news was positive or not.

There were multiple news items on some days. The news text content was merged together as the news data of the day. Based on the number of positive and negative news in a day, daily news was determined whether it was bullish, bearish or no impact. At the same time, there were also some days without news text content. This study used the emotional characteristic value of the previous day multiplied by a coefficient to represent the emotional characteristic value of this day since the emotional characteristics of each day were also affected by the news content of the previous day. This study took the value of this coefficient as 0.2.

2.2. Proposed Approach

In this study, the BERT model was utilized to extract text features, which were then fed into a Back Propagation neural network for emotion calculation, resulting in the final emotion classification of news. The model comprised three main layers: an input layer converting each word into vectors, a BERT layer employing the Transformer Encoder to extract features and fuse full-text semantics, and a BP neural network emotion calculation layer using the SoftMax activation function to determine the emotional characteristics of financial news titles. These emotion classifications were combined with stock trading data and input into a Transformer model to predict stock index fluctuations. The pipeline is shown in the Fig. 1.

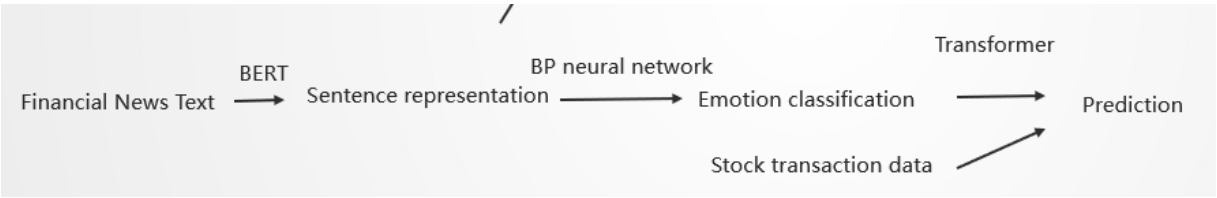

Figure 1. The pipeline of the model

2.2.1. BERT.

The model first obtained the vector representation through BERT. The pre-training model in this study is the BERT-base-Chinese model which is more suitable for Chinese tasks. This study segmented the text content of financial news headlines about 300 Shanghai and Shenzhen stocks into words. BERT performed word segmentation based on individual Chinese characters to associate each word with its index in the vocabulary. The maximum text length in this study was set as 148 characters. Text longer than 148 characters was truncated and text less than 148 characters was filled with 0. classification (CLS) and separator (SEP) identifiers were added at the beginning and the end of the input text. The BERT model input was obtained by adding the word vector, segment vector and position vector. Two pre-training tasks were carried out by the BERT model after input, which were Mask Language Model (MLM) and Next Sentence Prediction (NSP). The model must be capable of predicting obscure words, comprehending contextual information, and creating appropriate representations in the pre-training task for MLM. The NSP pre-training task requires the model to predict if two sentences are related and comprehend the logical connection between them. After the

pre-training tasks were completed, BERT was fine-tuned according to subsequent tasks. This study performed an emotional analysis task and used the output information of CLS as the input of the next layer of the network. And then the text data with emotional annotations was used to fine-tune. Finally, a model with higher accuracy in predicting stock rises and falls based on financial news was trained.

2.2.2. BP neural network.

This study input the feature vector CLS output by BERT into the BP neural network and used the SoftMax function to predict emotion classification results. Finally, the model would output the probability that the data was positive or negative based on each piece of news text data. The BP network is a feedforward neural network with multiple layers that is trained using the error back propagation algorithm. The neural network that is most widely used is this one. The BP network adds one or more layers of neurons between the input layer and the output layer. The term hidden units is also used to refer to these neurons. Their state may have an effect on the relationship between input and output, even though they have no direct connection with the outside world. Each layer is capable of having multiple nodes. BP neural network calculates using forward and reverse calculations. The forward propagation process involves processing the input pattern layer by layer from the input layer to the hidden unit layer, and then transferring it to the output layer. The state of each layer of neurons is only a factor in the state of the next layer of neurons. If the output layer does not provide the expected output, it will turn to reverse propagation and return the error signal along the original connection path. By modifying the weights of each neuron, the error signal is minimized.

2.2.3. Transformer model.

This study used the Transformer model to predict the rise and fall of stock indexes in the next trading day because the Transformer model performed well in processing time series data. The core of the Transformer model's application in time series laid in its self-attention mechanism, which enabled it to effectively capture long-term dependencies in time series data. Through parallel processing capabilities and positional encoding, the Transformer model not only improved processing efficiency but also ensured chronological accuracy. The Transformer network consisted of multiple layers, as shown in the Fig. 2. The first layer was the sequence Transformer encoding layer, which was composed of multiple Transformer units. There were 64 attention heads in each unit, each of which had a self-attention mechanism and positional encoding. The input was the combination of stock transaction data and the emotional analysis feature for a single day. The second layer was the Transformer decoding layer. The decoding layer also used the Transformer structure. A fully connected layer was used to pass the feature vector learned by the Transformer layer through and use the sigmoid activation function to output the probability of the stock price rising or falling.

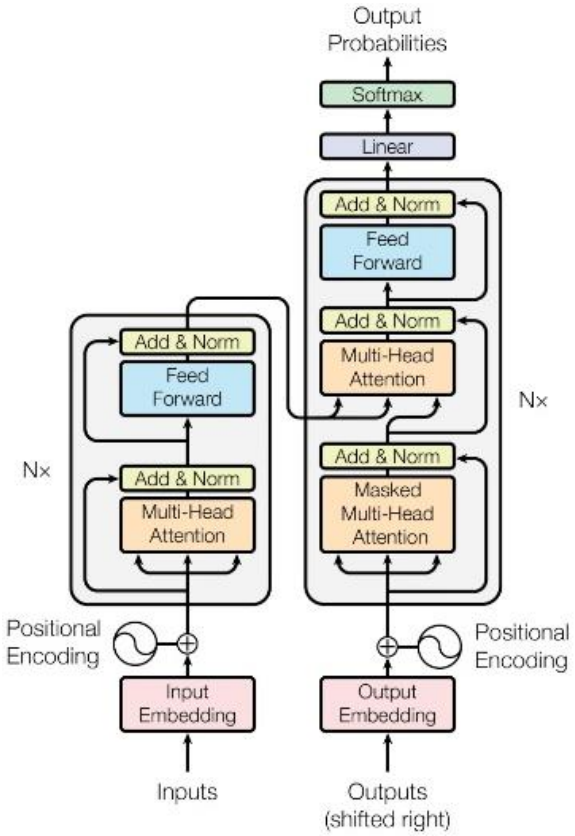

Figure 2. The Transformer-model architecture [13]

3. Result and Discussion

This study selected the currently commonly used stock index prediction models, including BP neural network, support vector machine (SVM) model and eXtreme Gradient Boosting (XGBOOST) model for comparison, to verify the effectiveness of the stock index prediction model. At present, the BP neural network is the most widely used neural network. It has great advantages in processing nonlinear and high-dimensional data. SVM model is a machine learning theory based on statistics. It performs better than other machine learning models when processing small samples, nonlinear and high-dimensional data. It is capable of overcoming the shortcomings of BP neural networks, like the local minimum and slow convergence speed, to a certain extent. XGBOOST is an algorithm that uses gradient-boosting decision trees to improve its performance. The characteristics of it include low computational complexity, fast running speed, and high accuracy, and can effectively prevent model overfitting.

This study used the accuracy, precision, recall and F1 values as evaluation indicators of prediction effect. The accuracy reflects the proportion of samples predicted correctly in the total samples; the precision reflects the proportion of samples predicted to rise that actually rise in the samples predicted to rise; the recall reflects the proportion of samples predicted to rise in the samples that actually rise; and the F1 value is the harmonic mean of the precision and recall.

Table 3. Experimental results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 value
BP	52.56	53.65	54.63	54.14
SVM	54.72	54.89	55.14	55.01
XGBOOST	56.46	57.23	57.61	57.42
BERT-transformer	61.4	60.59	61.39	60.99

Table 3 showed the experimental results. In conclusion, the prediction results of the BERT-Transformer model were better than those of the BP neural network, SVM and XGBOOST. The main reason may be that the above model only selects the market data of the previous trading day to predict the rise and fall of the next trading day, and the timeliness of stock information is obvious. Usually, the rise and fall of the trading day is closely related to the market trend and news sentiment characteristics of the previous few days. The BERT-Transformer model had long-term memory capabilities and was imported sentiment features. Therefore, it had better prediction effects on financial time series data. The stock price prediction model that integrates sentiment analysis features proposed in this paper achieved the best results. It proved the effectiveness of incorporating sentiment analysis into the stock index prediction model.

4. Conclusion

This study aimed to develop a novel BERT-Transformer model to enhance the accuracy of predicting stock market fluctuations, providing fresh research ideas and methods for financial researchers. The study involved collecting CSI 300 trading data and financial news text data, analyzing sentiment polarity characteristics within the financial news, and integrating structured and unstructured data to generate stock index predictions. The experimental results demonstrated that the BERT-Transformer model, with its long-term memory capabilities, excelled at predicting financial time series data, thereby proving the effectiveness of incorporating sentiment analysis into stock index prediction models. In the future, research will focus on automating the sentiment feature labeling process to reduce the time and effort required, aiming to streamline the model's application and enhance its efficiency. Additionally, further enhancements will explore the integration of real-time data and advanced sentiment analysis techniques to improve the predictive power and adaptability of the model in dynamic market conditions.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Hiew J.Z.G. Huang X. Mou H. et al. BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. 2019, arXiv preprint:1906.09024.

[2] Zhang L. Wang S. Liu B. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8 (4): e1253.

[3] Socher R. Pennington J. Huang E.H. et al. Semi-supervised recursive autoencoders for predicting sentiment distributions. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 151 - 161.

[4] Kalchbrenner N. Grefenstette E. Blunsom P. A convolutional neural network for modelling sentences. 2014, arXiv preprint:1404.2188.

[5] Kim Y. Convolutional neural networks for sentence classification. 2014, arXiv preprint:1408.5882.

[6] Graves A. Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18 (5-6): 602 - 610.

[7] Wang J. Yu L.C. Lai K.R. et al. Dimensional sentiment analysis using a regional CNN-LSTM model. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 2: 225 - 230.

[8] Mikolov T. Chen K. Corrado G. et al. Efficient estimation of word representations in vector space. 2013, arXiv preprint:1301.3781.

[9] Tang D. Wei F. Qin B. et al. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. Proceedings of Coling 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014: 172 - 182.

[10] Le Q.V. Mikolov T. Distributed Representations of Sentences and Documents. International Conference on Machine Learning, 2014: 1188 - 1196.

[11] Devlin J. Chang M.W. Lee K. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint:1810.04805.

[12] Kaggle dataset, “Crime Scene Investigation (CSI) 300 Index”, 2022, https://cn.investing.com.
[13] Vaswani A. Shazeer N. Parmar N. et al. Attention is all you need. 2017, arXiv preprint:1706.03762.