Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/69/20241474

Stock prediction using Bi-LSTM and RoBERTa sentiment analysis

Yiwen Wang

Maynooth International Engineering College, Fuzhou University, Fuzhou, 350108，China

YIWEN.WANG.2024@mumail.ie

Abstract. Stock price forecasting is a widespread research issue in the field of finance. Stock price prediction can give investors more accurate advice and yield higher returns. With the popularity of social media, more and more people choose to express their opinions, and more and more media outlets published news on social media. This paper mainly conducts stock analysis and prediction by presenting the RoBERTa-BiLSTM model with the Bahdanau attention mechanism. It consists of two existing models: the RoBERTa and bidirectional LSTM models. The experimental results show that this model is more accurate in predicting stock prices than the Multi-Layer Perception (MLP) and LSTM models. Inputs to the model include historical AAPL stock price data and sentiment data analysed by news and Twitter tweet sentiment. In this paper, the high-precision RoBERTa-BiLSTM model for stock prediction was proposed, which provides a solution for stock prediction. The practical applications and the typical problem of most studies about stock prediction called the “time-shifted problem”, are provided in this paper and also have research value.

Keywords: Stock price prediction, Bi-LSTM and RoBERTa model, sentiment Analysis

1. Introduction

Stock analysis and prediction lay the groundwork for informed decision-making, enabling investors to identify potential investment opportunities and assess risks. Research shows the company's stock price fluctuations are intrinsically linked to the sentiment conveyed in economic news coverage about the company [1]. Another study has unveiled a relationship between stock price volatility trends and sentiment expressed on online forums [2].

1.1. Sentiment Analysis

M. S. Neethu et al. conducted a text sentiment analysis using Support Vector Machines (SVM) and Naive Baye's (NB) [3]; Shiyang Liao et al. predicted the sentiment of Twitter posts using a convolutional neural network (CNN) [4]; R. Monika et al. used the Long Short Memory Model (LSTM) to analyse the sentiment from text [5]. Bidirectional Encoder Representations from Transformers (BERT) is a language representation model developed by Google [6]. BERT has shown powerful performance in natural language processing tasks. Matheus et al. also used the BERT model for sentiment analysis [7]. The result shows that the accuracy of validation datasets is relatively high.

1.2. Stock Prediction

The Bahdanau Attention mechanism enhanced the ability to model long-term dependence by assigning different weights to inputs. Zhao et al. used the Bahdanau Attention mechanism and the LSTM model; the result shows that the model can predict the stock price accurately [8]. The bidirectional LSTM (Bi-LSTM) consists of a 2-layer LSTM and a 1-layer output merger layer. With the bi-directional structure of Bi-LSTM, the neural network can process both forward and backward information from the time series and better capture long-term dependencies. Arif et al. used Bi-LSTM for stock prediction [9], and the model can fit the stock curve precisely.

1.3. Research idea

The BERT model has been shown to be extremely good at multiple language processing tasks, and both the Bi-LSTM and Bahadanau attention mechanisms have performed exceptionally well in processing time-related data. Sentiment data has been demonstrated to be crucial in stock analysis and prediction. RoBERTa is an improved BERT-based model. Inspired by these ideas, this paper proposes a RoBERTa-based model for sentiment analysis and a Bahadanau attention mechanism-based Bi-LSTM model for stock analysis.

2. Methodology

Stock price data for AAPL (Apple Inc.) for 720 valid trading days from 1 January 2014 to 31 December 2015 from Yahoo! Finance is selected in the paper. All news related to AAPL in that interval is obtained from the New York Times, while Twitter is used as the social media data source. The experiment aims to take the news and Twitter posts and perform sentiment analysis using the RoBERTa-based model, aggregate the respective sentiment indicators, and then output the normalised data volume (news of the day and number of Twitter posts of the day). Historical stock price data and sentiment indicators predict the next day's stock price.

2.1. Fetch data

Historical AAPL stock data can be downloaded from Yahoo! Finance, and the official New York Times API is used to fetch historical news data. Twitter datasets are from StockNet [10], another research project about stock price prediction.

2.2. Spam filtering and dataset import

Using the pre-trained Twitter-spam-classifier model on Hugging Face. Twitter-spam-classifier is a BERT-based test classification model developed by Delphia. The accuracy of the model reaches 78.3%. In this paper, tweets will be moved if the confidence value of the spam is above 0.9.

2.3. Sentiment analysis

Using the Twitter-RoBERTa-base model for sentiment analysis [11]. It is a RoBERTa-based model trained on about 124 million tweets from January 2018 to December 2021. The pseudo-perplexity indicator for the Twitter-RoBERTa-base model reaches 4.489, which shows good performance in sentiment analysis. Pseudo-perplexity was an evaluation metric for masked language models introduced by Salazar et al. [12]. The output of the model contains a label and a confidence value. Then, using the DistilRoberta-financial-sentiment model from Hugging Face for news sentiment analysis. It is another RoBERTa-based sentiment analysis model used for news.It has shown 98.23% accuracy in the evaluation set. The output of the model includes a label and a confidence value. All models are pre-trained.

2.4. Consolidate all the data

Import all data into the same CSV file for input. The file contains the date, stock close price, news sentiment, number of news, forum post sentiment, and its number.

2.5. Building Neural Networks for Prediction and Comparison

2.5.1. Build the MLP model. The MLP model consists of a fully connected layer with a Dropout added to prevent overfitting. The model consists of three Dense layers and two Dropout layers.

2.5.2. Build the LSTM model. Figure 1 shows the structure of an LSTM cell. An LSTM cell consists of 3 gates: an input gate, an output gate, and a forget gate. The input gate decides what new information from the current input and previous state to update into the cell state, and the forget gate controls what information from the prior state to discard or keep selectively. The output gate decides what information from the current cell state to output as the final output for the current time step.

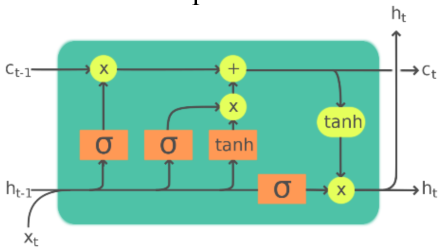

Figure 1. LSTM cell [13]

Table 1 shows the structure of the LSTM model. The LSTM model uses three LSTM layers: LSTM to get better information before and after the stock price, Dropout to prevent overfitting, and finally, a fully connected layer whose output is the stock price data for the next day. In this paper, a three-layer LSTM with dropouts are constructed.

Table 1. Structure of the LSTM model

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 12, 70)	20,160
dropout (Dropout)	(None, 12, 70)	0
lstm_1 (LSTM)	(None, 12, 50)	24,200
Dropout_1 (Dropout)	(None, 12, 50)	0
lstm_2 (LSTM)	(None, 20)	5,680
dropout_2 (Dropout)	(None, 20)	0
dense (Dense)	(None, 1)	21

2.6. Build the RoBERTa-BiLSTM model.

Figure 2 shows the Bi-LSTM structure. The Bi-LSTM model is an improved version of the regular LSTM model. A Bi-LSTM model consists of two connected LSTM layers: forward LSTM and backward LSTM. The forward LSTM captures context from the beginning to the end of the sequence, and the backward LSTM fetches data in opposite directions. At each time step, the Bi-LSTM combines the forward hidden state and backward hidden state in the complete context, which leads to superior performance in processing time-related problems.

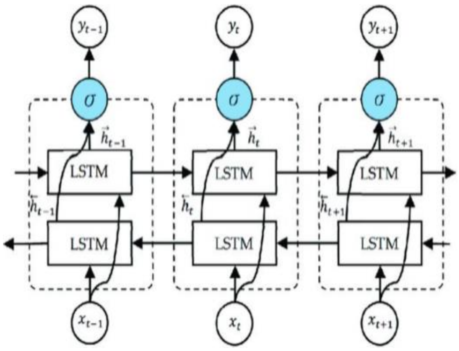

Figure 2. Structure of Bi-LSTM [14]

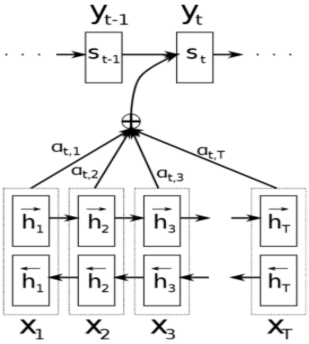

Figure 3. Bahdanau Attention Decoder [15]

Figure 3 shows the basic structure of the Bahdanau Attention Decoder. The Bahdanau Attention mechanism has been widely used in the machine translation field. At each time step, the decoder computes an alignment distribution based on the previous output and all the hidden states of the encoder, pointing out which positions in the encoder are most accurate. The decoder then uses the weighted encoder's hidden states to generate the current output.In this paper, the Bi-LSTM layer and the Bahdanau Attention mechanism. Table 2 shows the structure of the RoBERTa-BiLSTM model. In this paper, Bi-LSTM layers and Bahdanau attention mechanisms are constructed. Since the two are parallel in the RoBERTa-Bi-LSTM model, a concrete layer is introduced for connecting features. The final output contains only the next day's stock price prediction.

Table 2. Structure of the RoBERTa-BiLSTM model

Layer (type)	Output Shape	Param #	Connected to
input	(None, 11, 1)	0	-
bidirectional (Bi_LSTM)	(None, 11, 140)	40,320	input[0][0]
dense (Dense)	(None, 11, 70)	9,870	bidirectional[0][0]
attention (BahdanauAttention)	[(None, 11, 70), (None, None, 11,1)]	10,011	dense[0][0], dense[0][0]
get_item_1 (GetItem)	(None, 70)	0	attention[0][0]
concatenate(Concatenate)	(None, 210)	0	dense[0][0], dense[0][0]
dense_4 (Dense)	(None, 30)	6,330	concatenate[0][0]
dense_5 (Dense)	(None, 1)	31	dense_4[0][0]

3. Experiment Analysis

3.1. Performance evaluation metric

Error analysis indicators, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), are used as evaluation metrics.

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|F_i - A_i| \tag{1}$$

$$\boldsymbol{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|\% \tag{2}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - A_i)^2} \tag{3}$$

Where $\mathrm{A_i}$ is the actual value and $\mathrm{F_i}$ is the forecast value of stock prices.

3.2. Experiment Result

The results are shown in Table 3. The metrics reveal that the RoBERTa-BiLSTM model proposed in this paper has a loss reduction of 4.7141% in MAE metrics and 6.3073% in RMSE compared to the

LSTM model, and it has a loss reduction of 19.9408% in MAE metrics and 16.1876% in RMSE compared to the MLP model.

Table 3. The prediction result by three different models

Model	MAE	MAPE	RMSE
MLP	0.44406121004728005	0.015471108236346497%	0.5273456531826183
LSTM	0.37310030349061585	0.013002687660902627%	0.47173522497384973
RoBERTa-BiLSTM	0.35551193649366447	0.012418852952586764%	0.44198126877628297

3.3. Error Analysis

From Figure 4, it can be concluded that the RoBERTa-BiLSTM proposed in this paper performs well in stock price prediction. However, the experimental results have some limitations. Potential factors contributing to the errors include the small coverage of the dataset, which only consists of 720 days of stock price data and its multivariate information for only one company, AAPL, the prediction accuracy of the spam classification model and the RoBERTa-based model itself, the under-labelling of the training set, and the hyper-parameter settings of the model itself. Follow-up work will likely focus on preprocessing the dataset and addressing the "time-shift problem" described below.

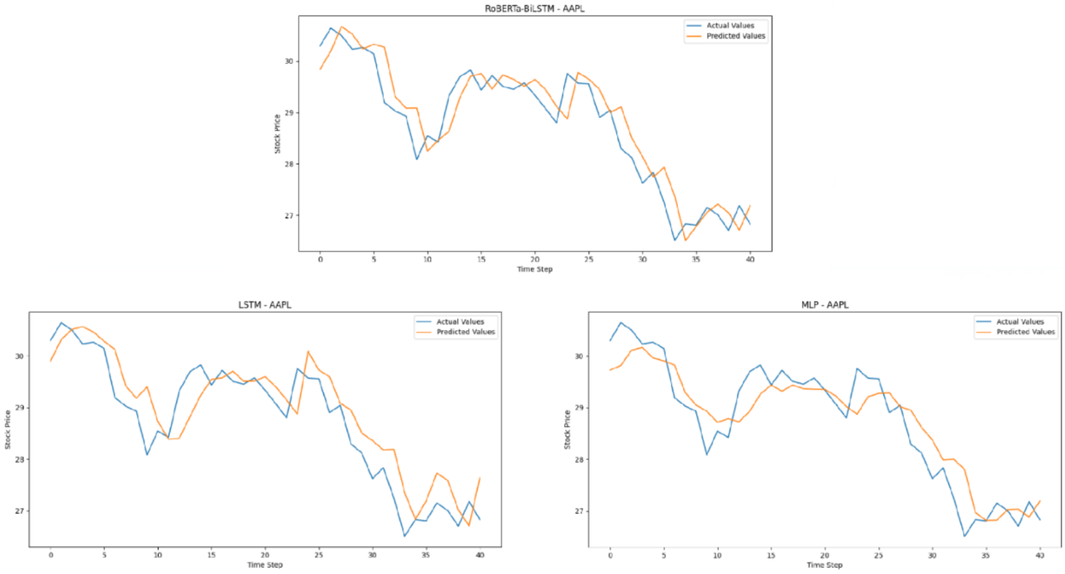

Figure 4. The prediction result by three different models

4. Discussion

4.1. Applications

4.1.1. StockTwits. StockTwits has gone live with a feature similar to stock forecasting. StockTwits provides users with two posting sentiment options: bearish and bullish. After that, the platform shows all users the number and sentiment of all the postings made on that stock topic for a certain period to help them make better decisions about investing.

4.1.2. Stock Prediction Websites. Some researchers have constructed stock price prediction websites. For example, Sonali et al. built a stock price prediction website using linear regression [16]. This paper proposes that local pre-trained predictive models can be brought online for further investment decisions by other investors.

4.2. Potential Challenges

In this paper, results predicted by models are compared to the previous day's data, which found that there was still a high degree of overlap. This means that the model remains insensitive to data from sentiment analysis. In this paper, this phenomenon is defined as the “time-shift problem”. This phenomenon is also reflected in other researchers' reports. Figure 5 shows the time-shift problem in this experiment.

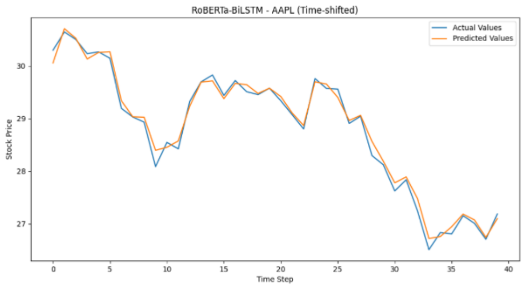

Figure 5. The time-shifted problem

If the volume of news and Twitter posts is added as input to the model, the "time-shifted problem" is mitigated. However, higher errors are produced than in the experimental results. Figure 6 shows the result of the RoBERTa-BiLSTM prediction after adding volume features of the news and tweets.

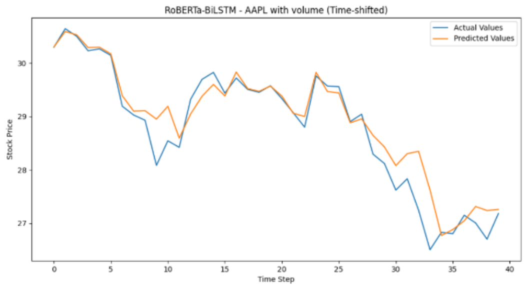

Figure 6. The RoBERTa-BiLSTM with the volume of data (time-shifted)

This property has yet to be investigated; here are some potential factors: the dataset has less data, does not cover a broad enough range of factors, and the accuracy of the work of the sentiment analysis model needs to be improved. The “time-shifted problem” evaluation metric has yet to be proposed.

5. Conclusion

This paper proposes an attention-based RoBERTa-BiLSTM modelling framework for stock analysis and prediction, and the importance of internal factors (historical stock data) and external factors (news sentiment and volume, tweets) that affect stock prices are assessed. Therefore, a deep learning method, RoBERTa-BiLSTM, is proposed based on multiple data sources for analyzing and predicting stock price data. The technique uses various data sources as indicators for stock analysis and prediction. In the experiment, AAPL stock price data is used for analysis and prediction and compared with other common MLP and LSTM models for stock analysis and prediction. The experiment results show that the RoBERTa-BiLSTM model has better prediction results. It is of great practical significance that investors can make better investment decisions through stock prediction results. However, this paper still has limitations due to a lack of time and research capability. It finds the common "time-shift problem" in stock forecasting, which needs to be solved by further research. The paper also explored the practical application of the model, which can provide advice to a broader range of investors by setting up a stock prediction website open to the public. This paper also has some shortcomings; the dataset is still relatively small, using only 720 active trading days for analysis, and due to time issues, more data from social media and news was not considered. The direction of stock prices is related to the historical stock price data used in this paper and news sentiment, crowd sentiment, and their quantities. Other factors affecting stock prices must also be considered.

References

[1] Alanyali, M., Moat, H. & Preis, T. Quantifying the Relationship Between Financial News and the Stock Market. Sci Rep 3, 3578 (2013). https://doi.org/10.1038/srep03578.

[2] D. D. Wu, L. Zheng and D. L. Olson. A Decision Support Approach for Online Stock Forum Sentiment Analysis, in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 44, no. 8, pp. 1077-1087, Aug. 2014, doi: 10.1109/TSMC.2013.2295353.

[3] M. S. Neethu and R. Rajasree. Sentiment analysis in twitter using machine learning techniques, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 2013, pp. 1-5. doi: 10.1109/ICCCNT.2013.6726818.

[4] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, Zixue Cheng. CNN for situations understanding based on sentiment analysis of twitter data, Procedia Computer Science, Volume 111, 2017, pp. 376-381, https://doi.org/10.1016/j.procs.2017.06.037.

[5] R. Monika, S. Deivalakshmi and B. Janet. Sentiment Analysis of US Airlines Tweets Using LSTM/RNN, 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 92-95, doi: 10.1109/IACC48062.2019.8971592.

[6] Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[7] M. G. Sousa, K. Sakiyama, L. d. S. Rodrigues, P. H. Moraes, E. R. Fernandes and E. T. Matsubara, BERT for Stock Market Sentiment Analysis, 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019, pp. 1597-1601, doi: 10.1109/ICTAI.2019.00231.

[8] Zhao R., Deng Y., Dredze M., Verma A., Rosenberg D., Stent A. Visual attention model for cross-sectional stock return prediction and end-to-end multimodal market representation learning. InThe Thirty-Second International Flairs Conference, 2019 May 19.

[9] Sunny M.A., Maswood M.M., Alharbi A.G. Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. In2020 2nd novel intelligent and leading emerging sciences conference (NILES), IEEE. 2020 Oct 24, pp. 87-92.

[10] Gupta U., Bhattacharjee V., Bishnu P.S. StockNet—GRU based stock index prediction. Expert Systems with Applications. 2022 Nov 30;207:117986.

[11] Loureiro D., Barbieri F., Neves L., Anke L.E, Camacho-Collados J. TimeLMs: Diachronic language models from Twitter. arXiv preprint arXiv:2202.03829. 2022 Feb 8.

[12] Salazar J., Liang D., Nguyen T.Q., Kirchhoff K. Masked language model scoring. 2019 Oct 31. arXiv preprint arXiv:1910.14659.

[13] Chevalier G. LARNN: linear attention recurrent neural network. arXiv preprint arXiv:1808.05578. 2018 Aug 16.

[14] Li Y.H., Harfiya L.N., Purwandari K., Lin Y.D. Real-time cuffless continuous blood pressure estimation using deep learning model. Sensors. 2020 Sep 30;20(19):5606.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[16] Stock Price Prediction Website Using Linear Regression - A Machine Learning Algorithm (Sonali Antad, Saloni Khandelwal, Anushka Khandelwal, Rohan Khandare, Prathamesh Khandave, Dhawal Khangar, Raj Khanke), ITM Web Conf. 56 05016, 2023. DOI: 10.1051/itmconf/20235605016.