NEU-Stock: Stock market prediction based on financial news*

Trang Tran[0000−0003−3370−6272], Nguyen Ngoc Long[0000−0002−6979−4473],
Nguyen Son Tung[0000−0001−9244−7093], Nguyen Thu Thao[0000−0002−4026−7362],
PTH Tham[0000−0001−9669−6855], and Tuan Nguyen**[0000−0002−3616−5267]

National Economics University, Hanoi, Vietnam
{huyentrang201ciel,ngoclong1282001,209tungns,thaothu2742001}@gmail.com
{thamtkt, nttuan}@neu.edu.vn

Abstract. For a long period of time, forecasting future stock price movements has attracted the attention of not only investors but also researchers. In this research, we examined the influence of financial news on the prediction of the stock price of FPT Group. At first, we presented a method to extract information from financial article titles and classified them based on their impact on stock prices by using a model that has been trained with PhoBERT with an accuracy of 93%. Then, we proposed a NEU-Stock model to forecast the stock price of the following day using the LSTM-Attention model with past closed prices and the impact of news as variables. The results of the tests demonstrate that utilizing the NEU-Stock model produces the best results with a high coefficient of determination R^2 and significant $RMSE$. The code is available at https://github.com/CielCiel1/NEU-Stock-Stock-market-prediction-based-on-financial-news.

Keywords: deep learning · stock prediction · news classification · LSTM · attention · PhoBERT

1 Introduction

Since the stock market is highly volatile and dynamic, forecasting is always a challenging task. Many methods have been proposed to forecast the stock market's future direction [1][2][3]. External factors such as financial news have an immediate positive or negative influence on stock values. For example, investors evaluate a business by its activities on its official website and financial news

* Copyright © by the paper's authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). In: N. D. Vo, O.-J. Lee, K.-H. N. Bui, H. G. Lim, H.-J. Jeon, P.-M. Nguyen, B. Q. Tuyen, J.-T. Kim, J. J. Jung, T. A. Vo (eds.): Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021), Da Nang, Viet Nam, 28-October-2021, published at http://ceur-ws.org

** Corresponding Author.

related to the company before they decide to buy that stock. However, the investors cannot completely assess such vast quantities of financial news data by themselves. Thus, investors naturally require a model that can anticipate stock prices.

Many prior studies predict the stock market using historical data[4][5]. However, the findings that those models offer are not particularly excellent since the stock market's volatility is highly impacted by unanticipated social network factors. Thus in this study, we used financial news to support the prediction of the stock market rather than only historical figures. Given news is in raw text, we introduced a PhoBERT[6]-trained model with the goal of classifying news-based emotions as negative, neutral or positive. The LSTM [7] model is then applied to combine the sentiment of the news with the historical stock price. Since the stock market fluctuations still contain a lot of noise, we decided to improve this model by using the attention mechanism to focus on the key information in the model. Finally, we proposed a NEU-Stock model that uses the LSTM-Attention model with historical closed prices and the impact of news as variables to predict the stock price of the next day. Once the model was trained on a data set including 1800 days of FPT stock price, our model produced the best results with an RMSE [8] error of 730.754 and the coefficient determinant R^2 [9] up to 0.933.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed model. Section 4 discusses the experiments and results, followed by a conclusion in Section 5.

2 Related works

Investment in the stock market is risky, but when arrived with discipline, it is one of the most accurate ways to earn large profits. Because accurate stock prediction external analysis increases investor profits, machine learning researchers are interested in this field. Wasiat Khan et al. [10] used algorithms to analyze social media and financial news data to determine the impact of this data on stock market prediction accuracy over the next ten days. Deep learning is used to achieve maximum prediction accuracy, and some classifiers are ensembled. Their experimental results show that social media and financial news have the highest prediction accuracies of 80.53% and 75.16%, respectively. Duc Duong et al. [11] proposed a model to predict the VN30 index trend based on stock news and the stock price of the VN30 index. They combined several methods such as delta TFIDF [12], sentiment dictionary, SVM [13], text mining to improve accuracy always above 60% (highest prediction accuracy is 90%). To develop the Arizona Financial Text System (AZFinText), Robert P.Schumaker and Hsinchun Chen [14] investigated the problem of discrete stock price prediction using a formulation of linguistic, financial, and statistical techniques (AZFinText). They discovered that stocks segmented by sectors were the most predictable in terms of closeness, with a Mean Squared Error (MSE) [15] score of 0.1954.

We proposed a new model that could predict stock prices after realizing that previous research only forecasted stock price trends. In addition, we provided a large and diverse data set to assist the model in making predictions with the lowest MAPE [16].

3 Method

3.1 Proposed model

In this paper, we presented the LSTM-Attention model for forecasting stock market closing prices based on the influence of news and historical prices. The method is built around two components: A stock news classification model and a prediction model based on LSTM and the attention mechanism. The complete model's operating procedure is as Figure 1: To begin, the model's input comprises the historical price of a FPT code and the title of financial news related to FPT Group, which is obtained from CafeF.vn. The stock price will be transformed into the value in the range (0,1) by taking each price minus the smallest price available in the data set and dividing the result by the distance between the smallest and highest price. This is to ensure the price distance is not too large among time intervals and simultaneously simplifies the computation process.

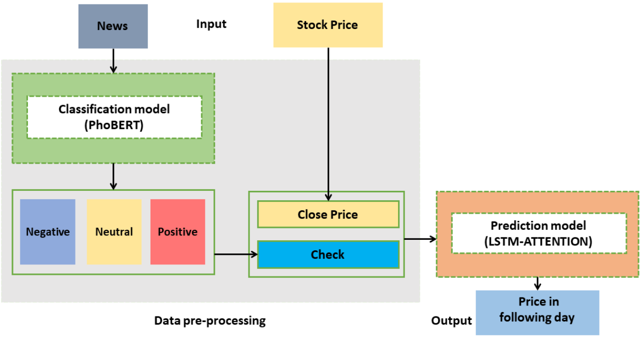

Fig. 1. The flowchart of NEU-Stock

Meanwhile, the titles are processed by the PhoBERT model. The algorithm will examine the sentiment impact of the title being broadcast by analyzing the content and categorizing it as [negative impact, neutral, or positive impact]. Following that, the model will count the number of articles categorize in each type of impact for each day. Then, based on that number, the model will decide whether that day is carrying a positive, negative, or neutral direction by taking the impact that has the highest number and represents as -1, 0 or 1 (i.e., -1 as negative, 0 as neutral, and 1 as positive direction). The outcome of this process, along with the scaled stock price, will be used as parameters in a model that uses LSTM-Attention [17] to train. The NEU-Stock model predicts the price of the following day using those parameters in a time series (in this case, 90 days) and then repeats the cycle day after day.

3.2 Stock price prediction model

Long short-term memory or LSTM neural network is powerful for modeling sequence data such as time series. It is a more advanced version of the RNN [18]. In comparison to RNN, LSTM consists of three gates to tackle the gradient vanishing problem, which is extensively used in time series modeling: forget gate, input gate, and output gate. Initially, data enters the forget gate in each neuron.

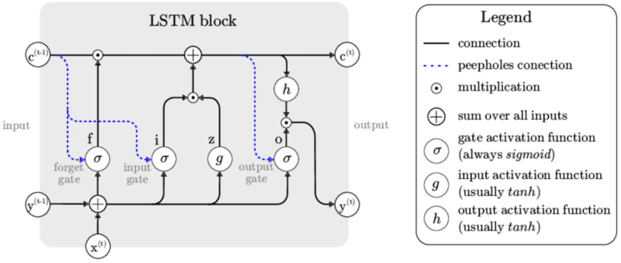

Fig. 2. LSTM architecture

The forget gate decides which input data is to be ignored so that the following neuron's update is not hampered. The input gate determines which data may be added in the second phase. The sigmoid and tanh function are used to process the preceding neuron's output and the local neuron's input to produce two outcomes. Then, depending on these two findings, it's decided which information has to be changed. For the output gate, the results will be stored. Finally, the output gate determines which of the input gate's results can be created. The findings of one neuron's output gate will be sent into the next neuron, and so on.

Attention. The attention mechanism is a part of a neural architecture that allows users to dynamically highlight significant characteristics of incoming data, which in NLP is generally a series of textual components. It can be applied to the raw input or its higher-level representation directly. The basic concept underlying attention is to compute a weight distribution on the input sequence, with larger values being assigned to more relevant items.

A query and a collection of key-value pairs are mapped to output by an attention method, with the query, keys, values, and output all being vectors. The result is a weighted sum of the values, with the weight allocated to each value determined by the query's compatibility function with the associated key. The following formula is used to determine attention's parameters:

$$\alpha_{ts} = \frac{exp(h_t, \overline{h_s})}{\sum_{s'=1}^{S} exp(score(h_t, \overline{h_s})} \qquad [Attention\ weights] \qquad (1)$$

$$c_t = \sum_{s} \alpha_{ts}\overline{h_s} \qquad [Context\ vector] \qquad (2)$$

$$a_t = f(c_t, h_t) = tanh(W_c[c_t; h_t]) \qquad [Attention\ vector] \qquad (3)$$

$$score(h_t, \overline{h_s}) = v_a^T tanh(W_1 h_t + W_2 \overline{h_S}) \qquad [Bahdanau's\ additive\ style[19]] \qquad (4)$$

3.3 PhoBERT

BERT, or Bidirectional Encoder Representations from Transformers, is an architecture for Language Representation published by Google [20] in early October 2018. The biggest advantage of BERT is the architecture designed to train the vector representing text language through two-dimensional context(from left to right and right to left).

However, it is not easy to apply BERT for Vietnamese because of the Vietnamese shortage of pre-training data. Almost publicly released monolingual and multi-lingual BERT-based language models are not aware of the difference between Vietnamese syllables and word tokens. This ambiguity comes from the fact that white space is also utilized to separate syllables that makeup words in Vietnamese. In March 2020, Dat Quoc Nguyen and Anh Tuan Nguyen from VinAI Research published pre-trained model PhoBERT. This is a monolingual pre-trained trainer, and the training is based on the design and approach of RoBERTa [22], which was introduced by Facebook in 2019 and is an improvement over the original BERT. PhoBERT is trained from about 20GB of data, including approximately 1GB of Vietnamese Wikipedia Corpus and 19GB remaining from Vietnamese News Corpus. Before proceeding to the BPE encoder [23], PhoBERT utilizes Rdrsegmenter of VnCoreNLP [24] to separate words for input data. The entire training process will be deployed on PyTorch [25].

4 Experimental results

4.1 Dataset

We developed two datasets, one to train the classification model, the other to train the stock price prediction model.

News classification dataset. To be able to use PhoBERT to evaluate and categorize the news' impact, we built a dataset that included 1000 titles of financial articles taken from CafeF.vn and labeled them into three groups [negative, neutral, or positive] based on expert advice. The dataset includes 187 articles with a negative impact, 248 articles with no impact, and 565 articles positively.

Stock price prediction dataset. Our dataset contains FPT stock prices and related articles. To get the stock price, we collected 1800 closing prices of

FPT stock from vn.investing.com for a period of seven years, between July 11, 2013, and September 24, 2020. Furthermore, we also crawled the news related to FPT Group from quality newspapers (e.g., CafeF.vn) in this period day by day. Then we classified the news's title by classification models and counted the number of positive, neutral, negative news each day. Finally, our dataset contains four main features as shown in Table 1.

Table 1. Feature description

Feature	Meaning
Price	The close price of FPT
Positive	The number of positive titles
Neutral	The number of neutral titles
Negative	The number of negative titles

4.2 Experimental results

The dataset is divided into two sets: training set with the first 1600 samples and testing set including the remaining 200 samples. We assessed LSTM and LSTM-Attention with news and without news performance based on the MAPE, RMSE, and R^2 metrics.

Table 2. Evaluate model's performance

	LSTM		LSTM-Attention	
	With news	Without news	With news **(NEU-Stock)**	Without news
MAPE	1.386	1.370	**1.338**	1.734
RMSE	747.777	748.348	**730.754**	854.851
R^2	0.932	0.932	**0.933**	0.911

It is observable from Table 2 that when we had the additional news feature, the model's performance is better. While the LSTM model with news is slightly better than the LSTM model without news in two out of three metrics except for MAPE, the LSTM-Attention model with news outperforms pure LSTM-Attention in all metrics. This is due to the fact that not all prices are equally important in predicting the price of the following day. As a result, we employ the additional Attention method to extract the crucial prices to forecast the price for the upcoming day. The NEU-Stock model achieved the best results amongst all the models, with RMSE = 730.754, and R^2 = 0.933, respectively. Because the news from many days ago still can affect the current stock price, the LSTM-Attention model can concentrate on important news and obtains the best prediction. The prediction of NEU-Stock on the test set is illustrated in Figure 3.

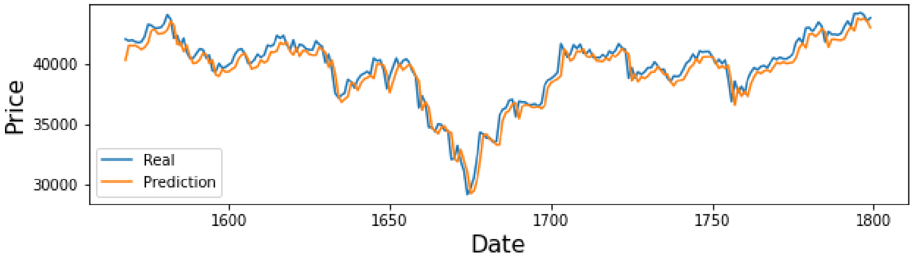

Fig. 3. The prediction of NEU-Stock model in test set

5 Conclusion

Stock prices do not fluctuate naturally; instead, they are influenced by a variety of external factors such as market situations, a company's upcoming plans, or the business's growth, etc. And all of these aspects are demonstrated clearly, objectively, and rapidly in prestigious financial articles. So in hypothesis, news plays an important role in predicting stock price trends. To clarify this, we collected articles related to FPT from CafeF.vn, which is a reputable website covering financial news and the stock market, then we categorized these news headlines using our PhoBERT-trained model with a current accuracy of up to 93 percent. Then we combined the outcome of the PhoBERT model with the stock price to formed the NEU-Stock model. The performance of the NEU-Stock outstood other techniques, according to tests performed on a dataset consisting of FPT's stock prices over 1800 days and news trendline in these essential aspects: closer forecast closing prices and greater accuracy of classification for bullish and bearish. Thus, it supports the belief that news has a significant impact on stock price forecasting. Therefore, the proposed approach in this paper is designed to assist investors in making the best decision possible. Furthermore, we believe that by applying and combining other factors that affect stock value growth and fall, we will be able to analyze and construct more practical models in the near future.

References

1. Pramod B S1, Mallikarjuna Shastry P.: Stock Price Prediction Using LSTM. Test Engineering and Management. 83. 5246-5251 (2020)
2. Sadia, K., Sharma, A., Paul, A., Sarmisthapadhi, S., Sanyal: Stock Market Prediction Using Machine Learning Algorithms. International Journal of Engineering and Advanced Technology (IJEAT). 2249–8958 (2019).

3. Budiharto, W.: Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). Journal of Big Data. 8, (2021).
4. Roondiwala, M., Patel, H., Varma, S.: Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR) ISSN. 6, 2319–7064 (2015).
5. Tang, J., Chen, X.: Stock Market Prediction Based on Historic Prices and News Titles. Proceedings of the 2018 International Conference on Machine Learning Technologies - ICMLT '18. (2018).
6. Nguyen, D.Q., Nguyen, A.T.: PhoBERT: Pre-trained language models for Vietnamese. (2020).
7. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation. 9, 1735–1780 (1997).
8. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. International Journal of Forecasting. 679–688 (2006).
9. Wright, S.: Correlation and Causation. Journal of Agricultural Research, 20, 557-585.(1921)
10. Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H., Alfakeeh, A.S.: Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing. (2020).
11. Duong, D., Nguyen, T., Dang, M.: Stock Market Prediction using Financial News Articles on Ho Chi Minh Stock Exchange. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication - IMCOM '16. (2016)
12. Martineau, J., Finin, T.: Delta TFIDF: An Improved Feature Space for Sentiment Analysis. (2009).
13. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their Applications. 13, 18–28 (1998).
14. Schumaker, R.P., Chen, H.: A quantitative stock prediction system based on financial news. Information Processing & Management. 45, 571–583 (2009).
15. Fürnkranz, J., Chan, P.K., Craw, S., Sammut, C., Uther, W., Ratnaparkhi, A., Jin, X., Han, J., Yang, Y., Morik, K., Dorigo, M., Birattari, M., Stützle, T., Brazdil, P., Vilalta, R., Giraud-Carrier, C., Soares, C., Rissanen, J., Baxter, R.A., Bruha, I.: Mean Absolute Error. Encyclopedia of Machine Learning. 652–652 (2011).
16. Ahmar, A.S.: Forecast Error Calculation with Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). JINAV: Journal of Information and Visualization. 1, (2020).
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan N, Kaiser, L., Polosukhin, I.: Attention Is All You Need(2017)
18. Graves, A.: Generating Sequences With Recurrent Neural Networks(2013)
19. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate(2021)
20. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(2018)
21. Gelbukh, A.: Natural language processing(2005)
22. Liu, Yinhan, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach.
23. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. Association for Computational Linguistics (2016).
24. Vu, T., Quoc Nguyen, D., Nguyen, D., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. (2018).
25. Paszke, Adam, et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library.(2019)