International Journal of Economics and Business Research

ISSN online: 1756-9869 - ISSN print: 1756-9850
https://www.inderscience.com/ijebr

FinBERT and LSTM-based novel model for stock price prediction using technical indicators and financial news

Gourav Bathla, Sunil Gupta

DOI: 10.1504/IJEBR.2023.10044437

Article History:
Received: 14 September 2021
Accepted: 06 December 2021
Published online: 01 July 2024

FinBERT and LSTM-based novel model for stock price prediction using technical indicators and financial news

Gourav Bathla

Department of Computer Engineering and Applications,
GLA University,
Mathura, India
Email: gourav.bathla@gla.ac.in

Sunil Gupta*

Department of Cybernetics,
School of Computer Science and Engineering,
University of Petroleum and Energy Studies,
Dehradun, India
Email: s.gupta@ddn.upes.ac.in
*Corresponding author

Abstract: Stock price movement is highly nonlinear, volatile, and complex. Traditional machine learning techniques are employed by researchers for stock price prediction, but due to shallow architecture, adequate accuracy is not achieved. In this paper, recently introduced bidirectional encoder representations from transformers (BERT) and long short-term memory (LSTM) hybrid model is utilised for stock price prediction. BERT model is used for financial news sentiment analysis. The sentiment score is merged with technical indicators of stock prices. In our approach, FinBERT is used which is specifically trained on financial corpus. Stock market prices were highly fluctuated in March 2020 due to lockdown. Thus, it is essential to predict high variation which existing works have not experienced due to lack of availability of highly fluctuated dataset. In our approach, the effect of financial news on stock indexes is analysed. Experiment analysis validates that our proposed approach outperforms existing approaches significantly.

Keywords: financial market; stock market; deep learning; data analytics; BERT; long short-term memory; LSTM; MAPE.

Reference to this paper should be made as follows: Bathla, G. and Gupta, S. (2024) 'FinBERT and LSTM-based novel model for stock price prediction using technical indicators and financial news', *Int. J. Economics and Business Research*, Vol. 28, No. 1, pp.1–16.

Biographical notes: Gourav Bathla is working as an Associate Professor at GLA University, Mathura, India. He has 15 years of teaching experience. He has completed his PhD from Punjabi University, Punjab, India. He has completed M.E from Delhi College of Engineering, India. He is GATE qualified with All India Rank 59. He is an active researcher and published 25 research papers in reputed journals and ten research papers in international

conferences. He has published three patents. His areas of interest are big data, machine learning, deep learning, NLP, and programming languages. He is a reviewer of various journals and TPC member of various international conferences.

Sunil Gupta has over more than 19 years of experience in teaching and research in the field of computer science and engineering. He is working as a Professor in University of Petroleum and Energy Studies (UPES). He is an Associate Member, Computer Society of India, member for Computer Science Teacher Association, life member of International Association of Engineers, member of International Association of Computer Science and Information Technology, Member, Internet Society (ISOC), and IEEE Society.

1 Introduction

Financial market is highly complex, volatile and nonlinear. A lot of innovations are proposed by researchers in financial technology (FinTech) to improve analytics for better decision making. For example, machine learning has improved analytics by FinTech significantly (Leow et al., 2021). However, the limitation of traditional machine learning technique is that it cannot perform complex computations due to its shallow architecture. This is due to the fact that multiple hidden layers are not used in shallow architecture. Further, recurrence and multiple regression is not possible in shallow architecture. These limitations are addressed by advanced deep learning models such as BERT, long short-term memory (LSTM) and RNN. For instance, there is concept of forget gate in LSTM that decides about output in next layer based on relevance. Thus, there is need for deep neural networks which can perform analytics on complex nonlinear data. In this research work, BERT and LSTM hybrid model is employed for analysing time-series data with very high variations.

Stock prices were highly fluctuated in March 2020 due to pandemic. NIFTY was down by 40% in just 30 days. It was estimated by Goldman Sachs that S&P 500 had endured 18% decline in three months. The stock market experts have compared these trends with year 1929 Great Depression. It was necessary to predict stock prices even during this high variation so that investors could save their holdings. In existing research works, this kind of variation is not experienced as nonlinear data was not available. In our approach, it is validated that if advanced models such as BERT and LSTM are applied with efficient hyper-parameter tuning, then even high variations can be analysed effectively.

The prediction of stock prices was assumed as impossible in earlier research works (Fama, 1970). However, due to advancements in computing, stock prices can be predicted with adequate accuracy now (Khashei and Bijari, 2011; Hoseinzade and Haratizadeh, 2019). In current era of advanced computing, stock price prediction is considered as hybrid of computer science and finance (Rezaei et al., 2021). Fundamentals of the company is reflected in news articles (Hagenau et al., 2013). Financial reports by analysts, organisation disclosures, and authenticated news articles significantly influence stock prices. Stock prices fluctuate when information about company is out (Boguth et al., 2016; Seong and Nam, 2021). External news which is unstructured needs to be analysed to predict stock price (Li et al., 2014a; Bustos and Pomares-Quimbaya, 2020).

Social media, reviews, and news have significant effect on stock price variations (Thakkar and Chaudhari, 2021). Furthermore, information from various sources can be merged to include information fusion which extends information about stock. Stock prices are predicted based on time-series of stock price technical indicators and textual data requires an efficient solution (Li et al., 2020). Further, sentiments in news feeds also has effect on stock price (Bharathi and Geetha, 2017; Lasek and Lasek, 2015). Several NLP techniques on textual news are used for financial forecasting (Xing et al., 2018). During the unprecedented time of year 2020, news has impacted sentiments of investors. This news cannot be analysed manually. Sentiment analysis is significant in stock market (Sonkiya et al., 2021). In this research work, BERT is applied to analyse the sentiments of investors and combine with numerical prices. The primary objective of this research work is to analyse sentiments of year 2020 which is not covered in existing research works. Furthermore, emotions in terms of news are given more focus in this research work. The terms 'lockdown', 'market shut down' etc. are given very high weightage in this research work. Stock prices are predicted by structured data, but local and global events are not covered in structured data (Mahajan et al., 2008).

It is concluded in several research works that deep learning approaches provide better prediction as compared to traditional machine learning approaches (Jing et al., 2021). It is due to the fact that traditional technique is not able to learn dependencies due to its shallow architecture (Pang et al., 2020). Researchers have proposed deep neural network to learn long-term dependencies to provide adequate accuracy. In stock market, large-scale transactions data is generated every day which can be used in deep learning models for better prediction (Li et al., 2020).

The major contributions of this research work are as follows.

1 The unprecedented stock price variations during March 2020 is predicted by our proposed BERT and LSTM hybrid model based on news sentiments and technical indicators.

2 FinBERT is employed for sentiment analysis which provides better accuracy as compared to general corpus-based sentiment analysis.

3 The news articles from stock index are scraped and employed for predicting variations. This contribution overcomes the limitation of existing techniques.

The rest of the paper is structured as follows. In Section 2, related work on stock price prediction using technical indicators and financial news are surveyed. The proposed approach is elaborated in Section 3. Section 4 covers the experiments with discussion of results. Finally, Section 5 concludes the paper with future directions.

2 Related work

Several research works have focused on predicting stock prices using technical indicators only. It is non-trivial to predict stock prices using historical data only (Mohan et al., 2019). Existing research works have predicted stock price using either numerical or textual information only (Akita et al., 2016). Companies and stocks information is not limited to numerical data only (Li et al., 2018). Stock price prediction based on historical prices is not efficient due to certain events which change stock prices (Minh et al., 2018).

Vargas et al. (2018) have stated that news and technical indicators are important factors that effect stock prices variations. Stochastic, momentum, rate of change, relative strength, etc. technical indicators are used in this paper. Furthermore, 106,494 Reuter's news articles as dataset are used for experiment analysis. The authors have further mentioned that intraday price prediction is more important than monthly price prediction. It is observed that CNN is better for analysing sentiments from news and LSTM is better for time-series prediction. The hybrid model is composed of CNN and LSTM as SI-RCNN and LSTM as I-RNN are proposed by authors. Experiment analysis has proved that the proposed model provides reasonable profit as compared to other strategies. In Hao et al. (2021), stock price is predicted using twin fuzzy logic using SVM. It is observed by authors that high/low variations are not classified accurately by traditional approaches. 10% variation is considered as equivalent to 1% variation in stock prices. This is the reason that fuzzy logic is applied in classification. The accuracy of proposed approach is compared with traditional SVM and fuzzy SVM approach. Emotional, topic, and embedding factors are used in this paper. It is also estimated that proposed approach using emotional, topic and embedding factors are better than using these features individually. In Mohan et al. (2019), it is stated that historical data is not sufficient to predict stock prices. Financial news articles and time-series data are used to predict stock prices. The authors have stated that previous studies have not collected large-scale data which has resulted in less accuracy. In this paper, large-scale price and news data is collected for S&P 500 companies and applied on ARIMA and RNN. Furthermore, RNN-LSTM model is used for different strategies such as price only, price and text polarity. It is concluded that model using news articles and price performs better than other models. Stock market predictions are applied based on the relationship between companies in Nam and Seong (2019). It is stated in this paper that previous studies have focused on news articles from individual company. There is need to observe causal relationships between companies. The advantage of this approach is that even if no news is available for target firm, stock price can be predicted based on causal firms. In this paper, transfer entropy is used for Global Industry Classification Standard Sectors (GICS). Korean market dataset is used and F1, accuracy are used as evaluation metrics. In Hagenau et al. (2013), the significance of news in stock price prediction is stated. Feature extraction, selection and representation are described in detail. Authors have used SVM for classification and 2-gram approach is used for feature extraction. It is also stated in paper that 2-gram approach is better than 3-gram approach, therefore 2-gram approach is used. The main contribution of this paper is to enhance the text feature extraction and selection which results in overcoming of over-fitting issue. Dataset is collected from new sources DGAP and EuroAdhoc. Experiment analysis proves that adequate accuracy is achieved. In Gidofalvi and Elkan (2001), textual news data is used to predict class of future news articles. It is observed in this paper that previous studies have focused on numerical data. There is need to analyse text data efficiently. Naïve Bayes classifier is used to train news article and stock price movement and then testing data is analysed for new articles. Time period of 10 minutes is fixed for 127 stocks news articles. Summarisation at sentence level is applied on financial news articles in Li et al. (2015). Authors have stated that a lot of mixed information is present in complete news article text. There is need to summarise the relevant text from news articles. Summarisation and complete news articles based models are compared in this research work. Experiments are conducted on Hong Kong Stock Exchange index dataset and news articles from Finet. Self-present sentence relevance model selects key sentences to assign

high scores. Further, SVM is used for text classification. It is concluded that the proposed approach SPSR outperforms existing approaches and SPSR using WordNet performs the best as compared to other approaches.

It is observed from comprehensive literature survey that traditional approaches are used for sentiment analysis of news articles. Even though neural networks are employed in some techniques, but these neural networks are unidirectional. We have overcome this limitation by using BERT which uses bidirectional technique for sentiment analysis and provides improved results.

For example, 'Company XYZ Makes New Investment to prepare workforce for opportunities of tomorrow. It commits to providing jobs as entry into digital careers for 50,000 job seekers by 2022'. In this stock news, FinBERT will use next sentence prediction (NSP) which predicts that second sentence is in continuation of previous sentence. Further, bidirectional encoder works from start and end of sentence in parallel.

Further, researchers have used general corpus and embedding to train neural network. There is need to utilise financial corpus and embedding. In our research work, FinBERT is used which is trained on financial corpus. Existing techniques have not faced high variations like unprecedented stock market variations in March 2020. In our research work, the prediction during high variation is provided accurately using FinBERT and LSTM. In existing research works, sentiments of specific company is extracted and merged with numerical prices to predict stock prices. However, to the best of our knowledge, the sentiments of overall stock index are not extracted. The fluctuations in stock index effect decision making of investors significantly. There are various stock index such as S&P 500, NASDAQ, NSE, BSE, NIFTY, NIKKEI 225, etc. In this research work, news about various index are scraped from web and sentiments about index is computed which is merged with technical indicators to predict stock price.

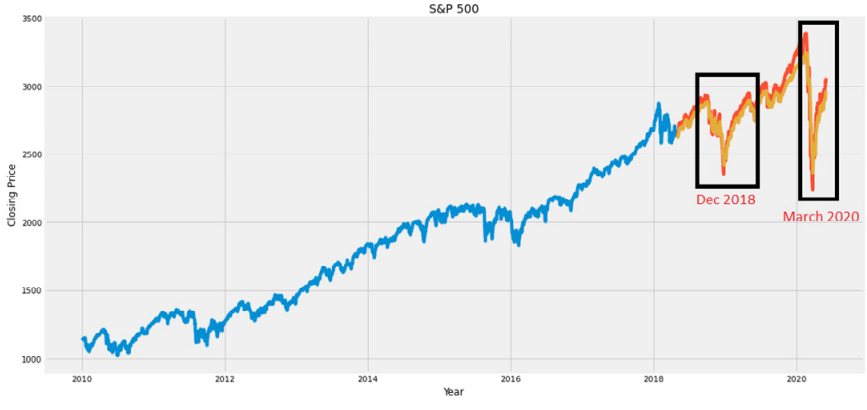
3 Proposed approach

The news about companies is shared on various platforms that results in Big unstructured data. It is very difficult for analyst to review news manually and predict stock prices based on sentiment of news. There is need for efficient text mining and NLP techniques to automate stock price prediction using news sentiments. Existing research works have employed bag-of-words (De Fortuny et al., 2014; Nam and Seong, 2019) and n-grams (Hagenau et al., 2013) approaches. The limitation of existing techniques is that complex computations for sentiment analysis are not performed effectively. Our proposed architecture is composed of two input layers. The first layer is textual news about particular company and the second layer is numerical technical indicators. In our proposed model, time window of 60 days and close price are used as technical indicator. Training of model is based on close price that indicates fluctuations due to previous 60 days price. Experiment analysis validates that time window of 60 days is suitable for training the model efficiently. There is no advantage of using other technical indicators such as average, standard deviation because LSTM model is able to learn long-term dependencies. Furthermore, our proposed model is robust as polarity of textual news from BERT is combined to predict stock price. NLP techniques such as stemming, lemmatisation, POS tagging and tokenisation are applied on textual news using NLTK

library to preprocess text of news. FinBERT is pre-trained on financial corpus and fine-tuned using news extracted using Finviz API.

In Figure 1, high variation in stock prices during December 2018 and March 2020 is depicted. It is clearly shown that stock prices have fluctuated significantly. Traditional machine learning or shallow neural network architecture cannot predict this high fluctuation. It is also observed that news has effect on stock prices in 1–5 days. It is not necessary that stock prices fluctuate on same day when news is out in press. It takes some time for the stock market to analyse and react to the news. This is the reason that the time window in our proposed approach is set to 2, 4, 6, 8, 10 days to observe the effect of news on different time frames. In recent approaches, stock price predictions based on news are analysed for same day only. The reality is that news takes some time to effect on stock price. Moreover, some terms are significant that effect stock prices immediately. In the same way, some news effect stock price slowly. In our proposed approach, news sentiments are computed based on specific terms.

Figure 1 High variation in stock prices during December 2018 and March 2020 (see online version for colours)



In our proposed work, textual news is fed into BERT model for sentiment analysis. Sentiment polarity which is output of BERT and historical prices are fed into LSTM model for learning time-series data and numerical polarity score. The significant contribution of our proposed approach is that specific terms that are outliers are assigned high weightage in polarity. The reason is that if these terms exist in news, stock prices highly fluctuate.

In Figure 2, our proposed model is depicted which is based on FinBERT and LSTM models. Textual news are fed into FinBERT model to extract polarity of news sentiments. This polarity and historical data are fed into LSTM model to predict stock prices. LSTM model is trained on polarity as well as numerical historical data that is quite better as compared to training based on numerical data only.

In our proposed model, LSTM is used as it is the most suitable model for time-series analysis. In existing research works, RNN was used, but the limitation of RNN is that gradient descent issue exists and long-term dependencies are not effective in RNN. In LSTM, time window of 60 days are used in our proposed model. The reason for using 60 days window is that training is better, i.e., price on x_{i+61} is dependent on x_i, x_{i+1},

x_{i+2},…….., x_{i+60} day price. Price on x_{i+62} is dependent on x_{i+1}, x_{i+2}, ……., x_{i+61} day price as depicted in Figure 3.

Figure 2 Proposed model using FinBERT, LSTM and historical data (see online version for colours)

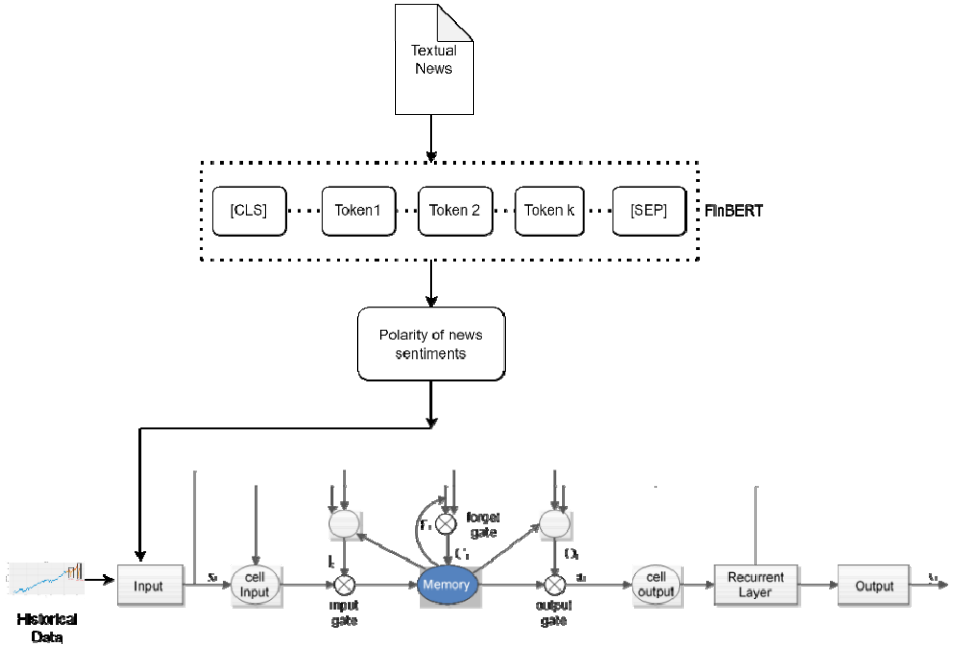


Figure 3 Stock prices of 60 days time window (see online version for colours)

X_i	X_{i+1}	X_{i+2}	X_{i+3}	.....	X_{i+60}	X_{61}			
	X_{i+1}	X_{i+2}	X_{i+3}	X_{i+4}	.....	X_{i+61}	X_{62}		
		X_{i+2}	X_{i+3}	X_{i+4}	X_{i+5}	.....	X_{i+62}	X_{63}	
			X_{i+3}	X_{i+4}	X_{i+5}	X_{i+6}	.....	X_{i+63}	X_{64}

3.1 BERT

Bidirectional encoder representations from transformers (BERT) is introduced by Google to develop models such as language inference, sentiment analysis etc. (Devlin et al., 2018). BERT can also be applied for machine translation and sentiment analysis. It is pre-trained on large unsupervised Wikipedia dataset. The sentiment analysis is efficiently applied in BERT. It uses transformers which are employed to analyse sentiments using encoder for input and decoder for output (Sonkiya et al., 2021). BERT has gained unprecedented attention in recent days (Leow et al., 2021).

The phases of BERT are as follows.

1 Pre-training: In this phase, context learning is achieved using unsupervised learning. In mask language model (MLM), mask of words is assigned. MLM is also used for left and right training of sentences (Hoang et al., 2019). In NSP, sequence of sentences is predicted.

2 Fine-tuning: In this phase, various NLP tasks can be solved using supervised learning. The parameters are fine-tuned during this phase. [CLS] is used for every sentence and [SEP] is used to separate different sentences.

Token, segment, and position embedding are used in BERT for input. In this paper, news sentences are assigned to sequence [[CLS]] x_1, x_2, ….., x_n [SEP]], where x_i is term i in news. [CLS] is token that is used for sentence classification and [SEP] is used for different sentences in NSP.

$$h = BERT(x) \tag{1}$$

where x is input of news terms. This input is fed into dense layers where softmax function is used.

$$L_i = soft\max\left(w_i x + b_i\right) \tag{2}$$

where w_i is weight and b_i is bias.

The sentiment analysis of specific domain stock-related news is different from sentiment analysis of common domains. In general terms, ‘happy’, ‘disappointed’, ‘delighted’ etc. are used as polarity, but in financial news terms such as ‘rise’, ‘down’, ‘jump’, ‘investment’, ‘buy’, ‘sell’ are used as polarity. Training using general corpus cannot be applied to finance-related sentiments. In our proposed approach, financial corpus is used for training purpose. FinBERT (Araci, 2019) is used for sentiment analysis in our approach. Financial corpus is used for training of FinBERT and fine-tuning is done for financial sentences classification. Furthermore, finance phrase bank (Malo et al., 2014) is used for sentiment analysis. Several financial dictionaries are used by researchers to train dataset. In Li et al. (2014b), Harvard and Loughran-McDonald financial sentiment dictionaries are used. However, in Loughran-McDonald financial sentiment dictionary is based on word counting and semantics of words are not analysed effectively (Araci, 2019). The main reason for using BERT for sentiment analysis is that context of sentences is considered effectively as compared to existing embedding. The sentiment class from BERT is merged with data frame having numerical prices. The merged dataframe is fed into LSTM for predicting stock prices using time-series analysis.

3.2 LSTM

LSTM model is used for long-term dependencies and time-series analysis. It overcomes the limitation of recurrent neural network (RNN). Input, forget and output gate are the main components of LSTM. Forget gate is used to filter output from previous layer and if value is 1, information is retained and if value is 0, information is forgotten.

Various components of LSTM are as follows.

1 Cell state: The information is saved in cells.

$$c_t = F_t * c_{t-1} + I_t * \overline{c}_t \tag{3}$$

where F_t is Forget gate state, c_{t-1} is previous cell state, I_t is Input gate state.

2 Input gate: Input is forwarded via this gate.

$$I_t = \sigma\left(w_{ix} x_t w_{ia} a_{t-1} + w_{ic} c_{t-1} + b_i\right) \tag{4}$$

where w_{ix} is weight of input x, x_t is input, a_{t-1} is previous activation function, c_{t-1} is previous cell state.

3 Forget gate: In this gate, only relevant information is saved.

$$O_t = \sigma\left(w_{ox} x_t + w_{oa} a_{t-1} + w_{oc} c_t + b_o\right) \tag{5}$$

where w_{ox} is weight of output x, x_t is input, a_{t-1} is previous activation function, c_t is current cell state.

4 Output gate: The output is forwarded to next layer.

$$F_t = \sigma\left(w_{fx} x_t + w_{fa} a_{t-1} + w_{fc} c_{t-1} + b_f\right) \tag{6}$$

where w_{fx} is weight, x_t is input, a_{t-1} is previous activation function, c_{t-1} is previous cell state.

4 Experiment analysis

In this section, sentiment analysis of financial news is analysed using BERT. The sentiment score is merged with dataframe of numerical prices. The combined sentiments and numerical prices are trained using LSTM which results in better prediction. In next subsections, experiment setup, dataset and discussion of results are presented.

$$MAPE = \frac{1}{n} \sum_{i=0}^{n} \left| \frac{y_{acti} - y_{predictedi}}{y_{acti}} \right| * 100 \tag{7}$$

MAPE score is calculated as difference of actual value and predicted value and divided by number of testing data.

4.1 Experiment setup

Experiments are conducted on Google Colab GPU. NLTK library is used for preprocessing such as stopwords removal, stemming, lemmatisation, POS tagging, etc. Keras 2.3.0, Pandas 1.2.3, and Numpy 1.20.1 libraries are used for our neural network architecture. This preprocessed text is fed into BERT for sentiment analysis. Sentiment score is merged with numerical prices and technical indicators. This merged dataframe is sent as input into LSTM. Training and testing in LSTM is done on the basis of time window of 60 days. Earlier, training was on the basis of numerical prices in traditional implementations. In our proposed approach, sentiment score is applied in synchronisation with numerical prices to enhance prediction. The duration from year 2012–2020 is used

for training purpose and year 2020–2021 is used for testing purpose. The reason for selecting stock prices from 2020–2021 year for testing is very high fluctuations in stock prices were observed in this duration.

BERT base model is used for experiment analysis which contains 12 encoder layers, 12 bi-directional self-attention heads, and 768 hidden units. In LSTM, 60 units are used in first layer with dropout 0.2, in next layer, 80 units are used with same dropout. In next layers 80 and 120 units are used with same dropout.

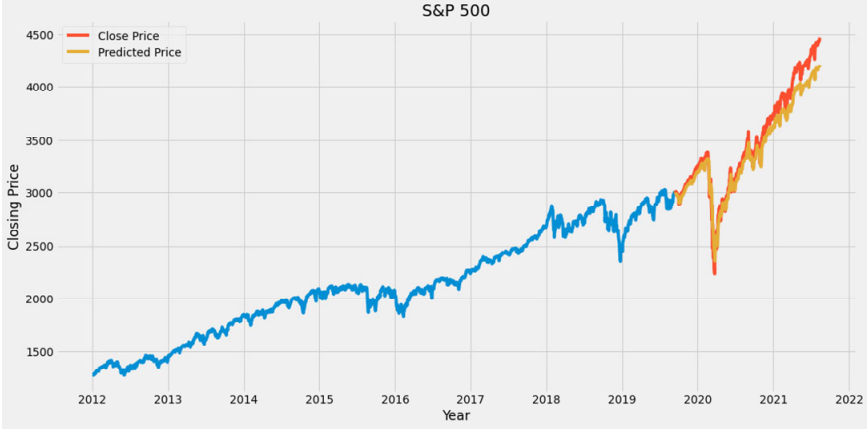
4.2 Dataset

In this research work, various stock index such as NSE, S&P 500, NYSE, BSE, Nikkei 225, NASDAQ are used. The news articles are scraped using Finviz API and merged with numerical prices. Yahoo finance API is used for extracting numerical prices. The numerical prices are extracted from year 2012 to 2021. The duration from year 2012–2020 are used for training purpose and year 2020–2021 is used for testing purpose. Time window for training is assigned to 60 days.

4.3 Results and discussion

Stock prices are predicted for various stock index and results are presented in this subsection. In Figures 4–9, the results of prediction for NSE, S&P 500, BSE, NASDAQ, NYSE, and Nikkei 225 are depicted. It is clearly shown during testing that even high variations are predicted with adequate accuracy.

Figure 4 Prediction of S&P 500 stock index (see online version for colours)



In Table 1, MAPE score of proposed model is compared with existing approaches. Support vector regression (SVR) is used for stock price prediction in Meesad and Rasel (2013). Extreme learning machine (ELM) is used in Mohanty et al. (2020). LSTM is used in Mehtab and Sen (2019). The results validate better performance of our approach as depicted in Figure 10.

Figure 5 Prediction of NSE stock index (see online version for colours)

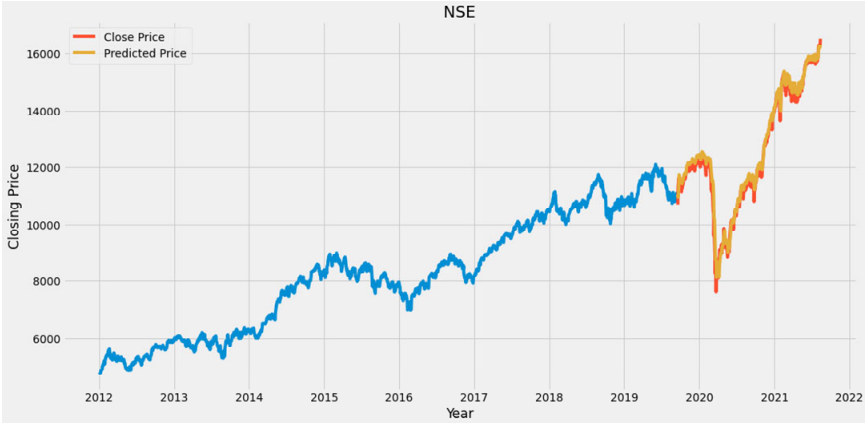


Figure 6 Prediction of BSE stock index (see online version for colours)

Figure 7 Prediction of NASDAQ stock index (see online version for colours)

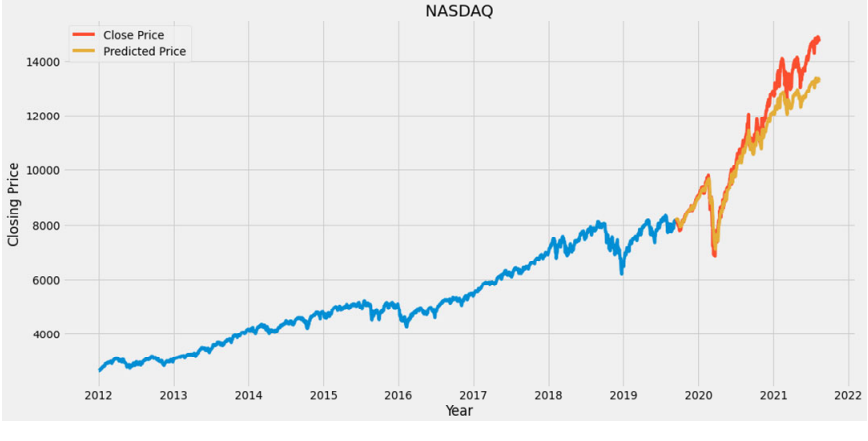


Figure 8 Prediction of NYSE stock index (see online version for colours)

Figure 9 Prediction of Nikkei 225 stock index (see online version for colours)

Table 1 MAPE scores of existing approaches and proposed model

Stocks	*Proposed model*	*SVR (Meesad and Rasel, 2013)*	*ELM (Mohanty et al., 2020)*	*Mehtab and Sen (2019)*
NSE	1.85	3.22	3.85	4.25
BSE	2.12	4.11	4.99	5.20
NASDAQ	3.11	5.16	5.75	6.44
NYSE	1.08	2.15	2.77	3.15
S&P 500	1.12	2.82	3.12	3.70
Nikkei 225	1.10	2.70	3.04	3.55

Figure 10 MAPE score chart of existing approaches and proposed model (see online version for colours)

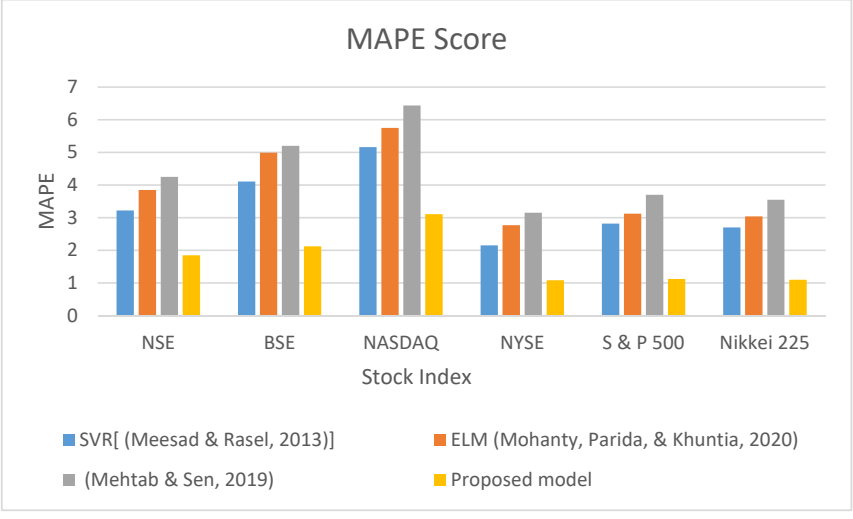


Table 2 MAPE scores using different LSTM units and epochs

Stock index	*LSTM units*	*Epochs*	*MAPE*
NSE	100	300	1.98
NSE	120	350	1.89
NSE	140	400	1.85
BSE	100	300	2.55
BSE	120	350	2.30
BSE	140	400	2.12
NASDAQ	100	300	3.70
NASDAQ	120	350	3.42
NASDAQ	140	400	3.11
NYSE	100	300	1.05
NYSE	120	350	0.92
NYSE	140	400	0.80
S&P 500	100	300	1.55
S&P 500	120	350	1.30
S&P 500	140	400	1.12
Nikkei 225	100	300	1.40
Nikkei 225	120	350	1.22
Nikkei 225	140	400	1.10

In Table 2, MAPE score is calculated using different LSTM units and epochs. It is clear that as LSTM units and epochs are increased, MAPE score is improved. The reason is that more training is done, thus, result is improved.

5 Conclusions and future directions

In this research work, the main limitation of existing works is addressed. Traditional approaches used only shallow architectures which do not provide adequate accuracy. In our approach, recently introduced BERT and LSTM hybrid model is employed to predict stock prices. In our proposed model, BERT is used for sentiment analysis, and LSTM is used for predicting time-series dataset. Our proposed hybrid model have predicted stock prices with very high variation which was experienced in March 2020. We have used FinBERT to enhance the accuracy as it is trained on financial corpus which is able to understand the sentiments hidden in financial and technical terms. Our approach is focused on observing the effect of financial news on S&P 500, NASDAQ, NYSE, BSE, NSE stock indexes. Experiment analysis has validated the better performance of our proposed model. MAPE score is improved as compared to recent approaches. In the future, training will be done on more financial corpus to train and test the model more effectively.

References

Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K. (2016) 'Deep learning for stock prediction using numerical and textual information', in *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, pp.1–6.

Araci, D. (2019) *Finbert: Financial Sentiment Analysis with Pre-Trained Language Models*, arXiv preprint arXiv:1908.10063.

Bharathi, S. and Geetha, A. (2017) 'Sentiment analysis for effective stock market prediction', *International Journal of Intelligent Engineering and Systems*, Vol. 10, No. 3, pp.146–154.

Boguth, O., Carlson, M., Fisher, A. and Simutin, M. (2016) 'Horizon effects in average returns: the role of slow information diffusion', *The Review of Financial Studies*, Vol. 29, No. 8, pp.2241–2281.

Bustos, O. and Pomares-Quimbaya, A. (2020) 'Stock market movement forecast: a systematic review', *Expert Systems with Applications*, Vol. 156, No. 1, p.113464.

De Fortuny, E.J., De Smedt, T., Martens, D. and Daelemans, W. (2014) 'Evaluating and understanding text-based stock price prediction models', *Information Processing & Management*, Vol. 50, No. 2, pp.426–441.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint arXiv:1810.04805.

Fama, E.F. (1970) 'Efficient capital markets: a review of theory and empirical work', *The Journal of Finance*, Vol. 25, No. 2, pp.383–417.

Gidofalvi, G. and Elkan, C. (2001) *Using News Articles to Predict Stock Price Movements*, Department of Computer Science and Engineering, University of California, San Diego.

Hagenau, M., Liebmann, M. and Neumann, D. (2013) 'Automated news reading: Stock price prediction based on financial news using context-capturing features', *Decision Support Systems*, Vol. 55, No. 3, pp.685–697.

Hao, P., Kung, C., Chang, C. and Ou, J. (2021) 'Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane', *Applied Soft Computing*, Vol. 98, No. 1, p.106806.

Hoang, M., Bihorac, O.A. and Rouces, J. (2019) 'Aspect-based sentiment analysis using BERT', in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Hoseinzade, E. and Haratizadeh, S. (2019) 'CNNpred: CNN-based stock market prediction using a diverse set of variables', *Expert Systems with Applications*, Vol. 129, No. 1, pp.273–285.

Jing, N., Wu, Z. and Wang, H. (2021) 'A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction', *Expert Systems with Applications*, Vol. 178, No. 1, p.115019.

Khashei, M. and Bijari, M. (2011) 'A novel hybridization of artificial neural networks and ARIMA models for time series forecasting', *Applied Soft Computing*, Vol. 11, No. 2, pp.2664–2675.

Lasek, M. and Lasek, J. (2015) 'Are stock markets driven more by sentiments than efficiency?', *Journal of Engineering, Project, and Production Management*, Vol. 6, No. 1, pp.53–62.

Leow, E.K.W., Nguyen, B.P. and Chua, M.C.H. (2021) 'Robo-advisor using genetic algorithm and BERT sentiments from tweets for hybrid portfolio optimisation', *Expert Systems with Applications*, Vol. 179, No. 1, p.115060.

Li, X., Huang, X., Deng, X. and Zhu, S. (2014a) 'Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information', *Neurocomputing*, Vol. 142, No. 1, pp.228–238.

Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014b) 'News impact on stock price return via sentiment analysis', *Knowledge-Based Systems*, Vol. 69, No. 1, pp.14–23.

Li, X., Wu, P. and Wang, W. (2020) 'Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong', *Information Processing & Management*, Vol. 57, No. 5, p.1022122020.

Li, X., Xie, H., Lau, R.Y., Wong, T.L. and Wang, F.L. (2018) 'Stock prediction via sentimental transfer learning', *IEEE Access*, Vol. 6, No. 1, pp.73110–73118.

Li, X., Xie, H., Song, Y., Zhu, S., Li, Q. and Wang, F.L. (2015) 'Does summarization help stock prediction? A news impact analysis', *IEEE Intelligent Systems*, Vol. 30, No. 3, pp.26–34.

Mahajan, A., Dey, L. and Haque, S. (2008) 'Mining financial news for major events and their impacts on the market', in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, Vol. 1, pp.423–426.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014) 'Good debt or bad debt: detecting semantic orientations in economic texts', *Journal of the Association for Information Science and Technology*, Vol. 65, No. 4, pp.782–796.

Meesad, P. and Rasel, R.I. (2013) 'Predicting stock market price using support vector regression', in *IEEE International Conference on Informatics, Electronics and Vision*.

Mehtab, S. and Sen, J. (2019) 'A robust predictive model for stock price prediction using deep learning and natural language processing', Available at SSRN 3502624.

Minh, D.L., Sadeghi-Niaraki, A., Huy, H., Min, K. and Moon, H. (2018) 'Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network', *IEEE Access*, Vol. 6, No. 1, pp.55392–55404.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D. (2019) 'Stock price prediction using news sentiment analysis', in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp.205–208.

Mohanty, D.K., Parida, A.K. and Khuntia, S.S. (2020) 'Financial market prediction under deep learning framework using auto encoder and kernel extreme learning machine', *Applied Soft Computing*, Vol. 99, No. 1, p.106898.

Nam, K. and Seong, N. (2019) 'Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market', *Decision Support Systems*, Vol. 117, No. 1, pp.100–112.

Pang, X., Zhou, Y., Wang, P., Lin, W. and Chang, V. (2020) 'An innovative neural network approach for stock market prediction', *The Journal of Supercomputing*, Vol. 76, No. 3, pp.2098–2118.

Rezaei, H., Faaljou, H. and Mansourfar, G. (2021) 'Stock price prediction using deep learning and frequency decomposition', *Expert Systems with Applications*, Vol. 169, No. 1, p.114332.

Seong, N. and Nam, K. (2021) 'Predicting stock movements based on financial news with segmentation', *Expert Systems with Applications*, Vol. 164, p.113988.

Sonkiya, P., Bajpai, V. and Bansal, A. (2021) *Stock Price Prediction using BERT and GAN*, arXiv preprint arXiv:2107.09055.

Thakkar, A. and Chaudhari, K. (2021) 'Fusion in stock market prediction: a decade survey on the necessity, recent developments, and potential future directions', *Information Fusion*, Vol. 65, No. 1, pp.95–107.

Vargas, M., dos Anjos, C., Bichara, G. and Evsukoff, A. (2018) 'Deep leaming for stock market prediction using technical indicators and financial news articles', in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.1–8.

Xing, F.Z., Cambria, E. and Welsch, R.E. (2018) 'Natural language based financial forecasting: a survey', *Artificial Intelligence Review*, Vol. 50, No. 1, pp.49–73.