International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Stock Price Prediction Using Ensemble Model and Sentiment Analysis

Anay Shah[1], Forum Sanjanwala[1], Anaya Jain[1], Radhika Gupta[1], Kiran Bhowmick[1], Nilesh Patil[1]

Submitted: 14/05/2024 **Revised**: 27/06/2024 **Accepted**: 07/07/2024

Abstract: The continued challenge in the world of finance is an accurate prediction of stock prices. This article investigates a trading system that uses machine learning algorithms to make recommendations on stocks to trade. This method integrates historical price analysis with sentiment examination in recent news articles. In this manner, the user is thus able to recognize patterns and predict future price movements using training models based on historical stock prices, turnover rates and other related indicators. Further, this study includes news articles' sentiments which tell how much news-driven sentiment affects the values of shares, and recommend the user whether to buy or sell or hold onto a particular stock. Performing standard metric tests on our approach and comparing its outcomes with traditional trading strategies, helps refine this study. By combining ratios like Simple Moving Average (SMA), Relative Strength Index (RSI) and fundamental analysis, gave ability to gauge the bullish or bearish behavior of a stock, aiding in providing a recommendation to the user based on previous data and further predictions. Through this research, evidence is provided within the algorithmic trading literature for machine learning and sentiment analysis that support datadriven stock recommendations. Implementing various machine learning models, this study concluded that an Ensemble technique using LSTM,RNN and GRU gave the best accuracy for the user with R^2 having value 0.9976.

Keywords: *LSTM,RNN,GRU,SVR,TextBlob,Sentiment Analysis,yfinance,*

1 Introduction

In today's world, the power of compounding holds the means to build a pathway to millionaire status if leveraged well or a considerable source of risk. Time being one of the biggest factors that plays a role in making one's investment portfolio, it is essential to invest early. However, while the stock market and its ways have tantalizing rewards, they are often beyond reach for young individuals who do not pursue their academic career in core finance or economics fields, such as engineering students. They are an excellent illustration of someone who may lack great skill analysis when it comes to selecting the appropriate stocks, solely due to lacking knowledge about the same. While recognizing this gap, there's a need to make investing easy and doable for all, to be able to explore the complex world of finance.

The purpose of this paper is to suggest an easy-to-use platform that empowers these people by requiring a login. The tool offers historical data from 2010 to the most recent trading day and focuses on the listed companies on BSE and Indian stocks. Machine learning approaches enable prediction of today's closing price for any stock one wishes for when models like Simple Recurrent Neural Network (RNN)[1][2][3], Gated Recurrent Unit(GRU)[4] , Long short term memory (LSTM)[5][2][3][4], Support Vector Regressor(SVR)[1] [6] and an ensemble of these models are used. Furthermore, sentiment analysis determines

whether the company has been evaluated favorably or unfavorably in recent news pieces about them and their products, so recommending whether or not to buy or sell.We use TextBlob[7][8] for sentiment analysis. In this manner, some stocks are suggested for purchase, while others should be sold right away, and some should only be held without being purchased or sold at all. They can then view their personal portfolio, which include all of their possessions and their currently bought shares, in a dedicated window of the interface. Therefore, in order to overcome the knowledge gap that prevents non-finance professionals from making wise stock market investments, this platform provides for the same.

2 Literature Review

Investing in stock markets is not a new investment idea at all rather the first stock market was formed in the year 1602 in Amsterdam. However with the risks involved in the stock market have always been high and with today's highly unpredictable and chaotic world an insight on how a stock will perform in the near future can be nothing less than a miracle. Stock price prediction is being done since a long time however there have always been few short comings. After studying [1] through which it was observed that stock price prediction was done using Artifical Neural Networks(ANN) and Random Forest(RF). It used the stock's 7, 14 and 21 days moving average, high minus low price, close minus open price and standard deviation for past 7 days to train the models. It used a stock's 10 year data for training the model. Overall ANN performed better compared to RF. Authors

[1]*Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.*
Email: -anayshah2705@gmail.com; forumsanjanwala@gmail.com; anayajain0309@gmail.com; guptaradhika213@gmail.com; kiran.bhowmick@djsce.ac.in; nilesh.p@djsce.ac.in;

suggested using deep learning models for better results. In this paper [2] the authors used both linear and non linear models in order to predict price of the stocks. Linear models used were AR,ARMA and ARIMA. While the non linear models were RNN,LSTM and CNN. They have used sliding window approach in order to make short term predictions of National Stock Exchange(NSE) stocks using one year data. Due to the nature of the price of the stocks the deep learning algorithms performed much better than the linear ones. Hence using deep learning models for prediction makes more sense even for long term predictions.In this paper [3] the authors employ models like RNN and LSTM to predict stock prices. The authors normalize the data to fit model and set the number of epochs as 10. The authors then run the model on stocks of Apple, Google and Tesla.The authors use the Markowitz Portfolio Optimization to understand how much money should be invested.The authors concluded that models like RNN-LSTM performed better than traditional machine learning models. In the paper [4] the authors first use LASSO to reduce the model deviation due to lack of independent variables. LASSO reduced the sum of square of residual under a constraint. Principal Component Analysis(PCA) is also used. PCA reduces the number of dimensions which helps in orthogonal transformation. The authors used data from Shanghai Composite Index from 2007 to 2021. They used LSTM and GRU to train this model. The authors use opening price, highest price, lowest price, trading volumes and other technical indicators to train the model. Both LSTM and GRU performed well while predicting the prices. In the paper [5] the author used Yahoo finance(yfinance) to get the data of Bitcoin and cryptocurrency prices and used LSTM in order to predict the prices. However LSTM did not perform as per the author's expectation. They felt that adding layers and maybe doing hyperparameter tuning will lead to better result while using LSTM, which we have considered while developing our model. In the paper [6] the authors have used stock index data like Bombay Stock Exchange(BSE), National Stock Exchange(NSE), S&P 500, NASDAQ, NYSE and Dow Jones Industrial Average to compare performance of LSTM and SVR. LSTM performed better than SVR. The author used 7 hidden layers, adam optimizer and 100 epochs for LSTM. The main difference is that the author uses stock indexes like NASDAQ or Dow Jones Industrial Average while our model is trained for individual stocks. Also author uses MAPE as a performance metric. We study the research paper [7] to understand how TextBlob is used for sentiment analysis. The author shows how TextBlob uses its in built library and dictionary in order to give a polarity and subjective score of each text unit. So TextBlob does a good job to analyse the sentiment along with how objective or subjective a text unit is by giving a score between -1 to 1 for polarity and 0 to 1 for subjective. However author highlights the drawbacks of TextBlob, which are it inability to detect emojis and detect sentiment when the text is in multiple languages. In the paper [8] we see how TextBlob can be an extremely handy tool in analyzing the sentiment of text in various situations.

Through this literature review significant insights were gained in order to narrow down on which machine learning and deep learning models we have to use in order to get better predictions. Across several papers LSTM was found to be a great model to predict the stock prices owing to the nature of the stock prices. Adding layers and making sure LSTM is optimized for the data is crucial for its performance. Also selecting the correct tool to do sentiment analysis is extremely important and TextBlob was an appropriate tool to do so for our application.

In addition to the extensive exploration of machine learning and deep learning models for stock price prediction, incorporating traditional econometric models such as the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model can provide unique insights into market dynamics. The GARCH model, introduced by Bollerslev in 1986, is adept at modeling and forecasting financial time series data characterized by volatility clustering—periods of high volatility followed by periods of low volatility. Mathematically, a GARCH(1,1) model can be expressed as:

$$r_t = \mu + \epsilon_t \epsilon t = \sigma t z t \, \sigma t 2 = \alpha 0 + \alpha 1 \epsilon 2 t - 1 + \beta 1 \sigma t 2 - 1$$

where r_t is the return at time t, μ is the mean return, ϵ_t is the error term, σ_t^2 is the conditional variance, z_t is a standard normal variable, and α_0, α_1, and β_1 are model parameters. This model explicitly accounts for time-varying volatility, making it an invaluable tool for understanding and predicting market risk.

By capturing the heteroskedasticity inherent in financial returns, GARCH models can complement neural network approaches. For instance, while LSTM models excel at capturing long-term dependencies and complex patterns in data, GARCH models provide precise volatility forecasts. This dual approach—combining GARCH for volatility estimation and LSTM for price prediction—leverages the strengths of both methodologies, offering a robust framework for stock price prediction and risk management. Integrating these models can enhance predictive accuracy and provide a comprehensive risk assessment, addressing the high unpredictability and chaotic nature of today's financial markets.

3 Methodology

a) Data Collection: For data collection Yahoo Finance(yfinance)[5][9] library was used to get the daily closing price for stock price prediction, news for sentiment analysis, volumes, historical prices etc. as it

is a reliable and trusted open source library which covers a wast variety of global financial markets. Data from 2010 to 2023 was used to train the models.

b) Data Preprocessing: Data collected from yfinance needs to be preprocessed in orderto make it usable for our models and to get more accurate results hence we employ the following techniques-

- Normalization - Normalization is basically done in order to make the data uniform across several features in order to make it better for predictions. MinMax scalar was employeed in order to normalize data between 0 to 1.
- Sequence Generation - Sequence generation is done on time series data by creating fixed length sequences which yield a input-output pair which is used to train models which will then be used for making predictions.
- Test-Train Split - The collected data was split into two parts, namely Test data which will be 80% which will be used to train the data and remaining 20% will be Test data which will be unseen data which will test the trained model.

c) Model Selection: For prediction of stock price 4 models were used, namely-

- Long Short-Term Memory(LSTM)
- Gated Recurrent Unit(GRU)
- Recurrent Neural Network(RNN)
- Support Vector Regression(SVR)
- Ensemble Model

d) Model Selection: A basic description of the models is as follows• Long Short-Term Memory(LSTM) - LSTM is a specialized version of recurrent neural networks (RNN) designed to solve the problem of vanishing gradient. This makes it possible to capture dependencies among data points that are far apart in time, which is why it works well for tasks such as predicting stock prices based on historical records.

- Gated Recurrent Unit (GRU) - GRUs are simpler architectures than LSTMs but still powerful enough to model temporal relationships. They have lower computational requirements and can be executed faster while preserving this ability through their integrated gating mechanisms.
- Recurrent Neural Networks - SimpleRNN lack the complex gate structure found in LSTMs or GRUs but act as basic building blocks for RNNs. They are less computationally expensive and good at capturing short-term dependencies over successive elements inside sequences – therefore, they should be used when trying out initial models or establishing baselines.
- Support Vector Regression (SVR) - SVRs represent a traditional machine learning technique for regression problems where support vectors are used to create a hyperplane that fits the training data best. It can handle nonlinear relationships quite well and also cope with large feature spaces.
- Ensemble Model - An ensemble model leverages the strengths of multiple machine learning algorithms to improve prediction accuracy. For stock price prediction, we implemented an ensemble model using three different types of recurrent neural networks: LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and SimpleRNN (Simple Recurrent Neural Network).To combine the predictions from these three models, we used a meta-model approach. A linear regression model was trained on the predictions from the LSTM, GRU, and SimpleRNN models. This meta-model learns to weigh the predictions from each base model to improve the overall prediction accuracy.

e) Performance Metrics: Two metrics were mainly used, namely • Mean Squared Error(MSE) - It is the average the squared value of difference between actual and predicted value. Lower MSE is an indicator of a better performing model and if it is 0 means it is perfectly predicting without any error which is extremely rare and difficult but possible. It is calculated using the formula :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (1)$$

$$\text{MSE percentage} = \left(\frac{\text{MSE}}{\text{Mean of the actual values}}\right) \times 100 \quad (2)$$

Coefficient of determination(R^2) - R^2 measures how well the model has fitted. It shows the ratio of variance in target feature that can be predicted from other features. R^2 is always between 0 and 1. 1 indicated model perfectly fits and R^2 value closer to 1 is preferred. It is calculated using the formula :

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Using this detailed methodology the prediction of the prices of stocks were performed. A flowchart explaining the methodology:

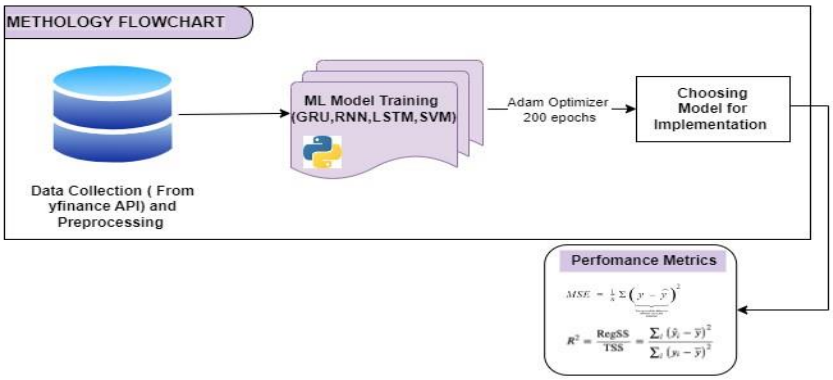

Fig. 1 Flowchart depicting the methodology

f) Sentiment Analysis: Latest news articles from yfinance library for any stock were used. Using that link and futher using, GeminiAPI[10] a summary of the news article was obtained. After getting the news article, lexicon based library in python was used, called TextBlob[7][8]. TextBlob assigns a polarity score to each unit of text between -1 to 1 where -1 is negative, 0 is neutral and 1 is positive and assigns a subjective score between 0 to 1 where 1 is subjective and 0 is objective to better understand the emotions in the text along with the sentiment. It does this based on pre-defined set of rules. Once the score is given for an article we repeat the process for 8 such articles and take the average. If the average is above 0 it recommends to buy the stock it analyzed however if it is below 0 it asks user to sell and if it is 0 then to hold the stock. A flowchart explaining the process is:

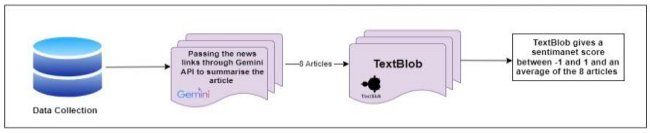

Fig. 2 Flowchart depicting the Sentiment Analysis methodology

4 Results

For testing, stocks of ADANIPORTS, RELIANCE, BAJAJFINANCE and UPL listed on National Stock Exchange(NSE), India were taken. Execution of LSTM,GRU,RNN,SVR and ensemble model was done and the results were as follows:

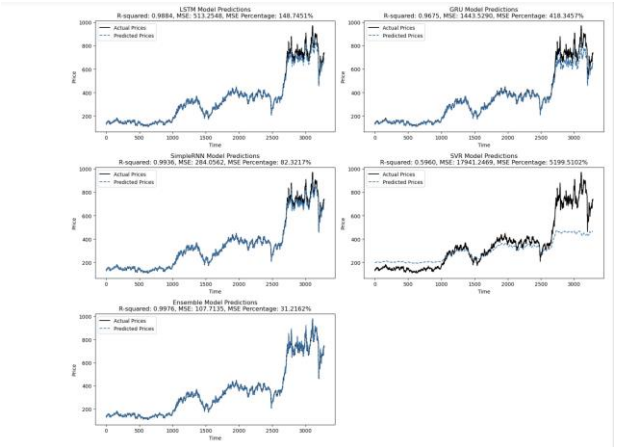

Fig. 3 Stock Prediction for Adani Ports

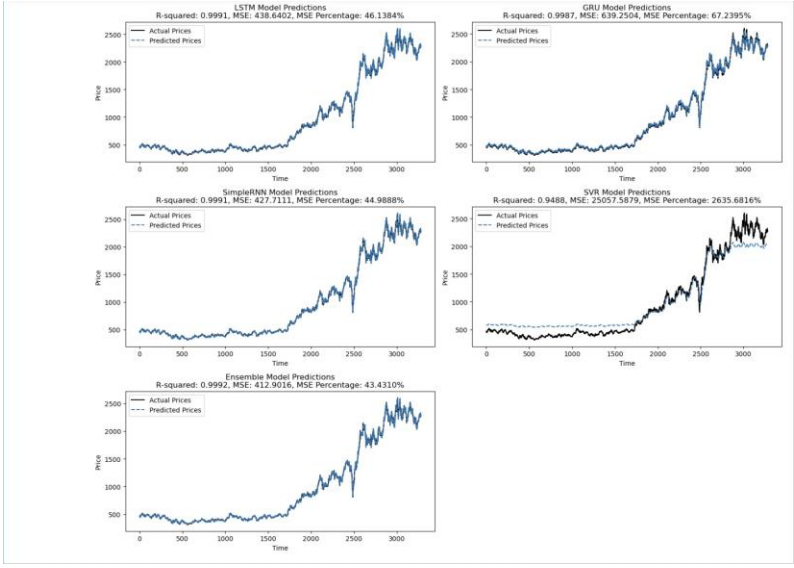

Fig. 4 Stock Prediction for Reliance

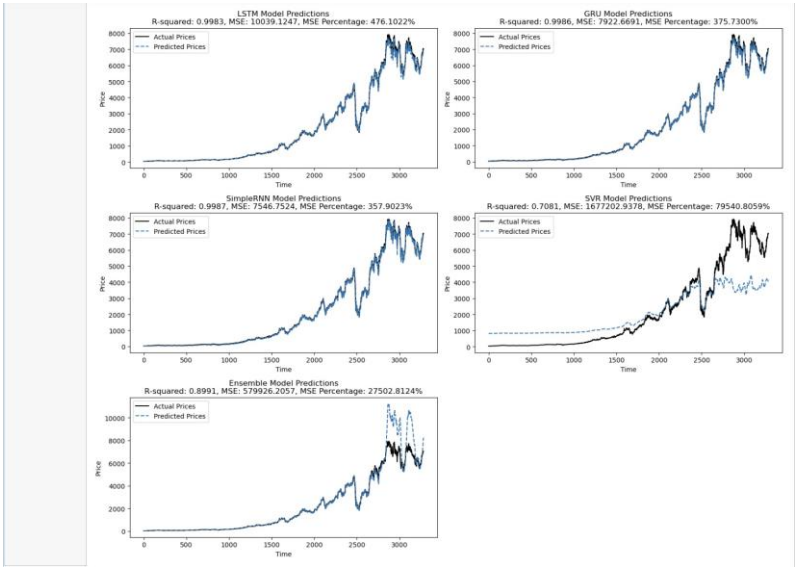

Fig. 5 Stock Prediction for Bajaj Finance

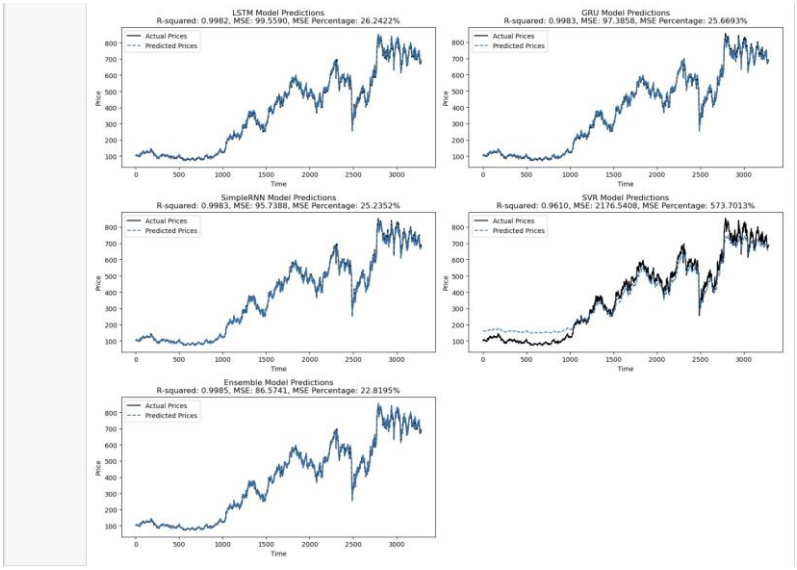

Fig. 6 Stock Prediction for UPL

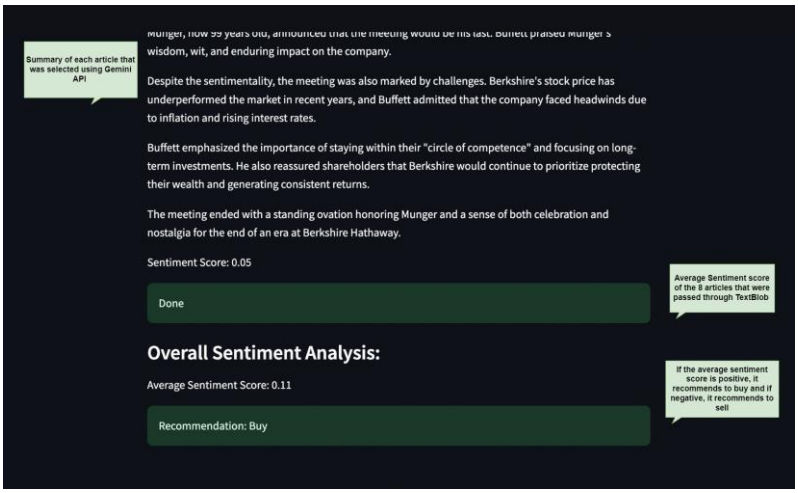

Fig. 7 Sentiment Analysis

5 Conclusion

Investing in the stock market is one of the surest ways to create wealth in the current world to multiply one's financial assets, although it is very complex and this usually

Table 1 MSE percentage for various stocks using multiple models

Stock Name	GRU	LSTM	RNN	SVR	Ensemble
Adani Ports	31.15%	32.48%	32.35%	546.30%	31.21%
Reliance	43.57%	56.02%	43.38%	3111.78%	43.43%
Bajaj Finance	255.21%	214.07%	224.96%	26684.83%	27502.81%
UPL	30.32%	21.37%	23.79%	508.56%	22.82%

Table 2 R^2 for various stocks using multiple models

Stock Name	GRU	LSTM	RNN	SVR	Ensemble
Adani Ports	0.997	0.997	0.997	0.961	0.997
Reliance	0.999	0.999	0.999	0.943	0.999
Bajaj Finance	0.999	0.999	0.999	0.906	0.899
UPL	0.997	0.998	0.998	0.964	0.998

discourages young people from participating. This study aimed at providing a solution to this problem by demonstrating how machine learning models together with sentiment analysis could help predict stock prices easily thus allowing those investing make informed decisions.

Having identified that many young adults are not financially literate, we probed into using Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) and Support Vector Regression (SVR) for stock price prediction. We collected historical data from platforms such as yfinance, fed them into the models whose architectures were carefully designed. In training these models, we used the R squared and Mean Squared Error loss function which ultimately led to highly accurate price predictions. This research indicated that the ensemble model performed the best with a high R square value and a low Mean Squared Error, proving to be the most efficient model amongst the ones tested, in predicting the accurate values for the user. These predictions are valuable insights for users so that they can choose their investments wisely, making investing easier. Additionally, using sentiment analysis to build this interface to further help the users in understanding the

standing of the company's stock recently, based on news articles published. Using tools like TextBlob to read between the lines of news articles and the hype surrounding a company on social media and understand how this will affect their stock price in the near future. This helps identify general public sentiment that could affect future markets in different directions.

It is essential to remember that it is the natural instability of financial markets that hinders any model's capability to give 100 percent accurate forecasts. Nevertheless, through these techniques' continuous improvement and merging them into easy-to-use interfaces, this study can equip young traders with valuable instruments— that ease investment procedures, fostered from an understanding approach to wealth generation, thereby enabling them to have more successful financial future.

6 Future Scope

These are some of the findings and areas for further detailed research that we uncovered while building this predictive model:

Stock Prediction using Sentiment Analysis and Historic Data: An enticing path forward would involve creating a unified model that seamlessly marries sentiment analysis with time series data and historical prices. The sentiment derived from news articles and social media is intertwined with temporal dependencies within price data; thus, painting a richer picture of market dynamics. By moving away from a myopic focus on historical price patterns alone — which is crowdsourced from public perception — we aim at better understanding industry specifics. This includes eschewing dependence on past price trends only based on traded volumes without questioning reasons behind these trends. - Ensemble Models: Using several ML models possessing different capabilities and weaknesses can potentially lead to more robust and accurate predictions through an ensemble model instead of a single model. - User Feedback: Design feedback loops that allow users to comment on the platform's recommendations and performance. This will enable continual adaptation and improvement based on user preferences.

References

[1] Vijh, M., Chandola, D., Tikkiwal, V.A., Kumar, A.: Stock closing price prediction using machine learning techniques. Procedia Computer Science **167**, 599–606 (2020) https://doi.org/10.1016/j.procs.2020.03.326 . International Conference on Computational Intelligence and Data Science

[2] Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K., Soman, K.P.: Stock price prediction using lstm, rnn and cnn-sliding window model. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1643–1647 (2017). https://doi.org/10.1109/ICACCI.2017. 8126078

[3] Pawar, K., Jalem, R.S., Tiwari, V.: Stock market price prediction using lstm rnn. In: Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, pp. 493–503 (2019). Springer

[4] Gao, Y., Wang, R., Zhou, E.: Stock prediction based on optimized lstm and gru models. Scientific Programming **2021**, 1–8 (2021)

[5] Ferdiansyah, F., Othman, S.H., Zahilah Raja Md Radzi, R., Stiawan, D., Sazaki, Y., Ependi, U.: A lstm-method for bitcoin price prediction: A case study yahoo finance stock market. In: 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), pp. 206–210 (2019). https://doi.org/10.1109/ ICECOS47637.2019.8984499

[6] Bathla, G.: Stock price prediction using lstm and svr. In: 2020 Sixth InternationalConference on Parallel, Distributed and Grid Computing (PDGC), pp. 211–214 (2020). IEEE

[7] Gujjar, J.P., Kumar, H.P.: Sentiment analysis: Textblob for decision making. Int. J. Sci. Res. Eng. Trends **7**(2), 1097–1099 (2021)

[8] Suanpang, P., Jamjuntr, P., Kaewyong, P.: Sentiment analysis with a textblob package implications for tourism. Journal of Management Information and Decision Sciences **24**, 1–9 (2021)

[9] Finance, Y.: Yahoo Finance. https://finance.yahoo.com/

[10] Gemini Trust Company, LLC: Gemini API Documentation. https://docs.gemini. com/rest-api/