MAY 2022

Gothapatana, Malipada,
Bhubaneswar, Odisha
751029

A PROJECT REPORT

on

Stock Market Analysis Using LSTM

Under the Guidance of

Prof. DEBANI PRASAD MISHRA

[DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING]

Submitted by

Prince Kumar [B320033]
Priyanka Rai [B320034]

TABLE OF CONTENT

ABSTRACT

Stock market forecasting is the process of estimating the long-term worth of a company's stock. This project explains how machine learning can be used to predict the stock market. The approach will function by feeding existing stock data into a random forest-style algorithm. We use a recurrent neural network, which recalls all of the data over time. By associating a loop that allows information to pass from one step to the next, LSTM networks save contextual information of inputs. Recurrent neural networks appear mystical because of these loops, to achieve the most accurate results since it produces an accurate outcome for future prediction. The prediction is compared using technologies such as Long short-term memory and the Random Forest regressor algorithm. Once the prediction result was much more accurate, this helps to make more profit of share price.

Keywords: Stock Market, LSTM, RNN (Recurrent Neural Network).

INTRODUCTION

Every day, a company's stock price fluctuates based on the trade of products and raw materials. The shares are provided by a corporation in order to improve the company's production. Predicting the value of a share is not an essay because it will fluctuate due to a variety of circumstances. Many investors aim to make a profit by investing their money in a firm. We have a trouble determining whether the company's share value will increase or drop by the end of the day. In this example, we're attempting to achieve perfect value through the use of technology. Technology advances at a breakneck pace.

While attempting to develop a machine learning model, there are numerous feathers to be acknowledged. The three primary values in a day's stock market are high, low, open, and close. Examine previous data to determine whether the value will rise or fall. To anticipate the value, many machine learning algorithms are employed, however for these types of time forecasting issues, we can use SVM and Recurrent Neural Network.For time series forecasting, SVM and LSTM offer unique features; try to implement them and improve accuracy.

We develop a model using LSTM and train it using historical data to predict 7 companies. We then examine the data based on the predictions. It can be difficult to anticipate the value in some circumstances because the value fluctuates during the day owing to other feathers. The cases the value goes wrong are when the company tries to implement new things or any political effects generally called a new feather.

BACKGROUND STUDY

Forecasting can be described as the analysis of historical data to anticipate some future occurrence or events. It covers a wide range of topics, including business and industry, economics, environmental science, and finance, to name a few. There are several types of forecasting issues.

- Short term forecasting (prediction for few seconds, minutes, days, weeks or months)
- Medium term forecasting (prediction for 1 to 2 years)
- Long term forecasting (prediction beyond 2years

The existing methods for stock price forecasting can be classified as follows

- Fundamental Analysis
- Technical Analysis
- Time Series Forecasting

Fundamental analysis is a sort of investment study in which a company's share value is calculated by examining its sales, earnings, profitability, and other economic aspects. This strategy is best for forecasting over a long period of time. For technical analysis, the past price of stocks is used to predict the future price. A moving average is a popular technical analysis algorithm. It's the unweighted average of the previous n data points. This strategy is best for making short-term forecasts. The analysis of time series data is the third way.It involves basically two classes of algorithms, they are

- Linear Models
- Non Linear Models

With the introduction of LSTM, the analysis of time dependent data become more efficient. These type of networks have the capability of holding past information.

PROPOSED ALGORITHM

LSTM networks are a type of neural network that are used in this research. They are deep and recurrent, and can hold a lot of information. Recurrent networks are different from standard feed-forward networks in that they include connections in both directions. Short-term memory and long-term memory are both affected by the way input travels in the brain. LSTMs were developed to improve performance by addressing the vanishing scaling problem faced by recursive networks while dealing with large data sequences, by keeping the error flow constant through special units called "gates" andAllows modification of weights as well as truncation of gradient when information about it is not necessary.

The **Input Gate** is the point where we pass the data into the network that we want to teach the model. Here there is another input from the output gate that is also called as a feedback signal that is the main one that was model learn what mistake it hades done and learn how to overcome from that loss. In these input gate, we use the sigmoid function and tanh function to combine the input and hidden value and get the output in the range -1 to 1.

The **Output Gate** is the last gate in the circuit we it gives the output value and it has an important duty that has to decide what data should be hidden for the next time. First, the previse hidden data and input passed to sigmoid function then pass the input value and sigmoid output multiply at tanh function that decides what should be stored for the next step.

The **Forget Gate** is the main gate that stores the previous information for farther references. The input from the previous data and the current data to the sigmoid function that gives the value 0 or 1. If we get 1 the forget gate store the data otherwise it will forget the present data.

$$\text{Update gate, } \Gamma_u = \sigma\left(W_u\left[h^{<t-1>}, x^t\right] + b_u\right)$$
$$\text{Forgate gate, } \Gamma_f = \sigma\left(W_f\left[h^{<t-1>}, x^t\right] + b_f\right)$$
$$\text{Output gate, } \Gamma_o = \sigma\left(W_o\left[h^{<t-1>}, x^t\right] + b_o\right)$$

Both the activation values and candidate values were employed in LSTM. As a result, the cell produces two outputs: one is the activation, and the other is the candidate value.

$$\text{Outputs } \begin{cases} c^{<t>} = \Gamma_u * c^{N<t>} + \Gamma_f * c^{<t-1>} \\ a^{<t>} = \Gamma_o * c^{<t>} \end{cases}$$

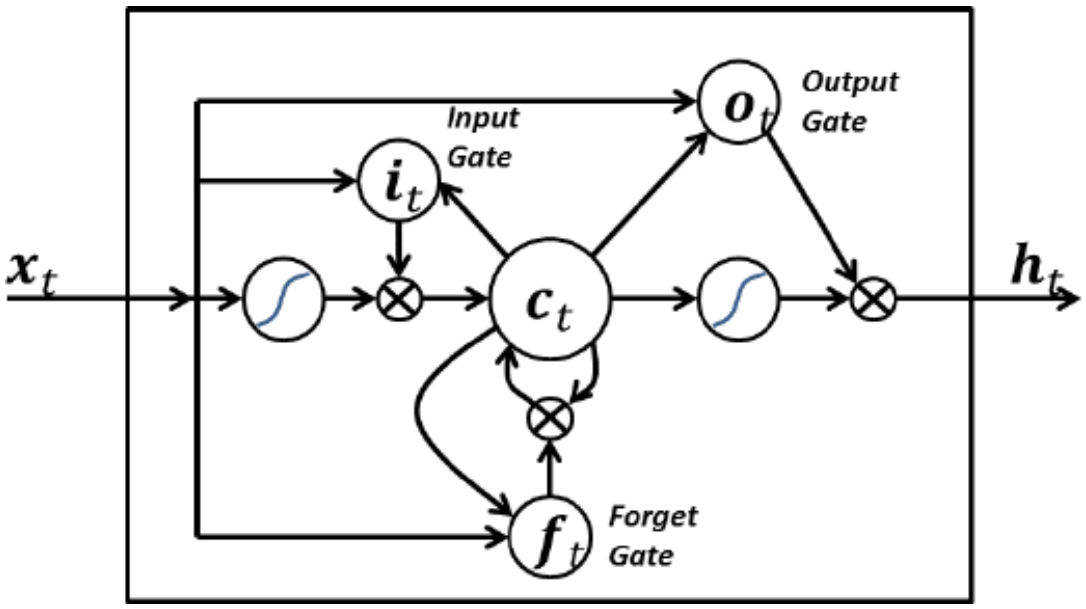

Fig. 4.1:
The repeating module in LSTM

METHODOLOGY

1. Download Share price Data

The raw data was gathered from yahoo finance, which is open to the public. This experiment is based on Google stock market data from August 19, 2004 to October 4, 2019. The input data is in numeric format. To predict future data, the data includes the stock's opening, high, low, and closing values on a daily basis. A total of 4170 days of data are used. In our research, we used Google colaboratory as a simulation environment, with a GPU, Ubuntu 18.04.3 LTS OS, and 12 GB RAM. CSV files are used to store the data.

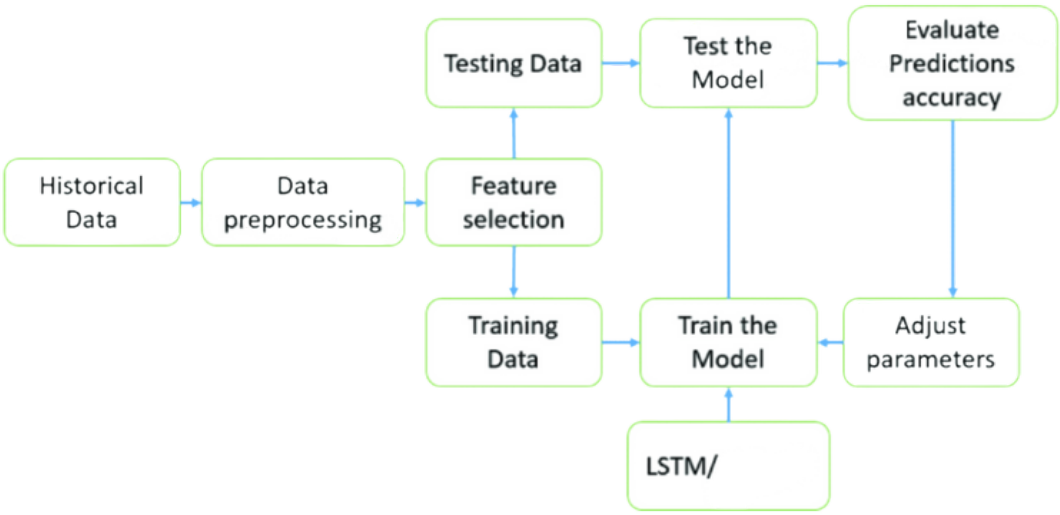

Fig. 5.1. The proposed architecture of stock price prediction

2. **Data Set**

The data set, as we all know, is the starting point for everything, thus it should contain enough data for the computer to learn about the problem. Scrap information available on the internet can be used to create or develop datasets. A dataset is a set of information. A data set usually consists of a single database table or statistical data matrix, with each column describing a specific variable and each row corresponding to a specific member of the data set in question. We keep all of our information in CSV files. A comma-separated values (CSV) file is a document that employs a comma to separate values in computing.CSV files are plain text files that include tabular data (numbers and text). A data record could be on each line of the file. Each record has one or more fields, which are separated by commas. The name for this file format comes from the use of the comma as a field separator. Our data is stored in a CSV file with values such as date, open, high, low, last, low, total trade, and turnover.

Date	High	Low	Open	Close	Volume	Adj Close
19-08-2004	52.08208	48.02803	50.05005	50.22022	44659000	50.22022
20-08-2004	54.59459	50.3003	50.55556	54.20921	22834300	54.20921
23-08-2004	56.79679	54.57958	55.43043	54.75475	18256100	54.75475
24-08-2004	55.85585	51.83684	55.67567	52.48749	15247300	52.48749
25-08-2004	54.05405	51.99199	52.53253	53.05305	9188600	53.05305
26-08-2004	54.02903	52.38238	52.52753	54.00901	7094800	54.00901
27-08-2004	54.36436	52.8979	54.1041	53.12813	6211700	53.12813
30-08-2004	52.7978	51.05606	52.69269	51.05606	5196700	51.05606
31-08-2004	51.90691	51.13113	51.2012	51.23624	4917800	51.23624
01-09-2004	51.53654	49.88488	51.4014	50.17517	9138200	50.17517
02-09-2004	51.23624	49.51952	49.64465	50.80581	15118600	50.80581
03-09-2004	50.92092	49.70971	50.52552	50.05505	5152400	50.05505
07-09-2004	51.05105	49.85485	50.55556	50.84084	5847500	50.84084
08-09-2004	51.56657	50.3003	50.42042	51.2012	4985600	51.2012
09-09-2004	51.40641	50.55055	51.31631	51.20621	4061700	51.20621
10-09-2004	53.33333	50.7007	50.85085	52.71772	8698800	52.71772

Fig. 5.2. Dataset of GOOGL

3. **Data Preprocessing**

A data set is a collection of feathers with N rows in which there are many values in various forms. There may be duplicate values or null values in a dataset, which can lead to loss inaccuracy and dependencies. Because data was gathered from many sources, numerous formats were used to represent a single value, such as gender, which is represented by M/F or Male/Female. Because the machine can only grasp 0 and 1, a 3-dimensional image should be reduced to a 2-dimensional format, such as data display, to avoid noisy data, null values, and an inappropriate format.
Panda's tabular data and OpenCV for images can be used to clean data.

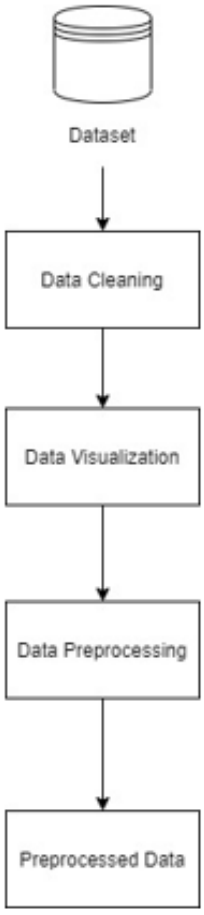

Fig. 5.3. Steps in data preprocessing.

4. **Feather Scaling**

The dataset contains numerical values that are steadily growing day by day. The closing value in the data set gradually increases from 1 to 2000. As in the case of the neural network target variable, a wide range of values can lead to big error gradient values, which can cause weight values to fluctuate dramatically, making the learning process unstable. As a result, we reduce the loss by scaling the dataset to a tiny size of (0,1) or (0,2). Scaling makes it simple for the model to store similar data for future reference.We scale all of the data from 0 to 1, then run the model with the scaled value as the input, reverting the scale to its original value.

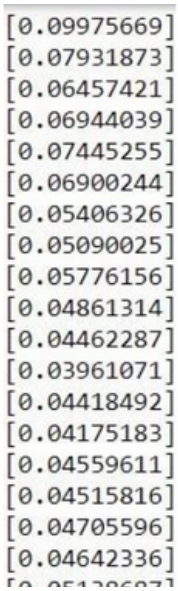

Fig. 5.4. Closed Values of the GOOGL after Scaling.

5. **Evaluation Metrics**

A variety of evaluation measures can be employed to estimate a prediction model's accuracy. One of these is the Root Mean Squared Error (RMSE). It is a widely used performance statistic for determining accuracy. yi and yi represent the original and forecasted values, respectively, while n denotes the total amount of data. The following is a description of the error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

RESULT AND ANALYSIS

To determine the closing value for the next day, we construct a fresh x list of 60 days values from today to the previous 60 days, which must be scaled and sent to the LSTM trained model to forecast the value that will give you the scaled value for the next day. To retrieve the original value, we reverse the scaled value. Because we're working with a time series, the LSTM neural network (Long short term memory) was chosen as the supervised learning technique. It's a recurrent neural network capable of classifying input data while taking prior instances into consideration.

The model was built using Google's TensorFlow and comprises of an LSTM input layer that will take both technical indicators and pricing data as input and will feed an output layer using sigmoid activation.

$$S(t) = \frac{1}{1 + e^{-t}}$$

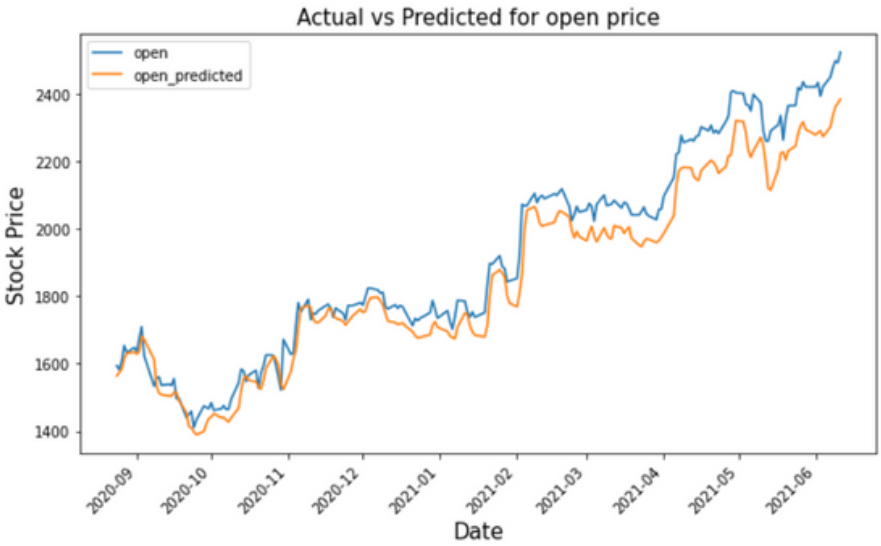

Fig. 6.1. Visualizing the predicted GOOGL open price.

Fig. 6.2. Visualizing the predicted GOOGL Close price.

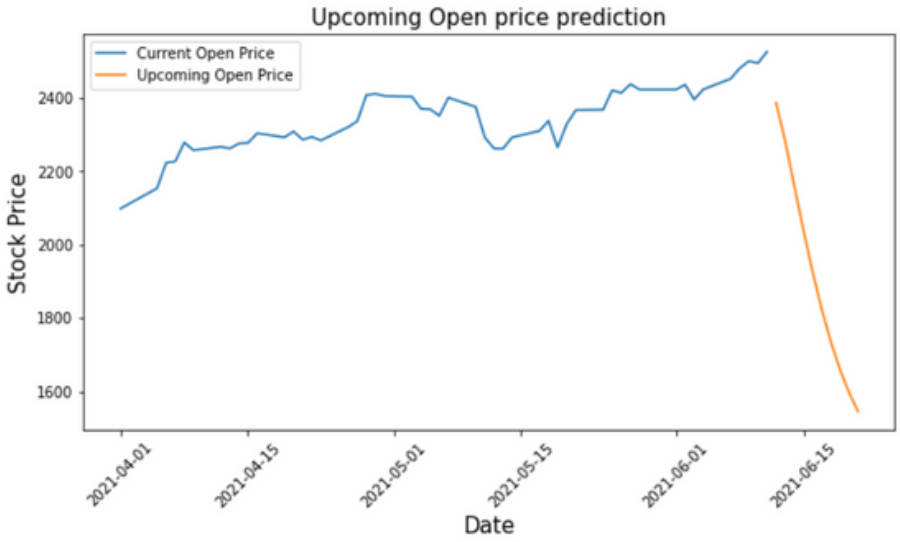

Fig. 6.3. Visualizing the GOOGL 10 days Open price prediction

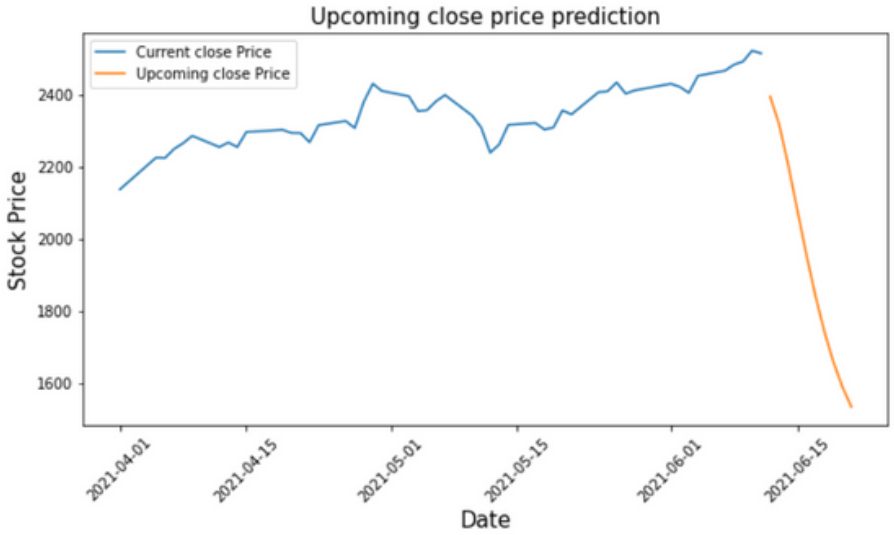

Fig. 6.4. Visualizing the GOOGL 10 days Close price prediction

```
rms=np.sqrt(np.mean(np.power((valid-closing_price),2)))
rms
```

```
11.772259608962642
```

We achieve the closest result for the model after running a lot of training studies in the LSTM model. The RMSE is substantially lower here, and the accuracy is much higher and overcomes the draw backs of the normal algorithm.

CONCLUSION & FUTURE WORK

Deep Learning algorithms have a considerable influence on modern technology, particularly in the development of distinct time series based prediction models, as shown in this study. They have the highest level of accuracy when it comes to stock price prediction when compared to other regression models.

With a few exceptions, we can see that the model proposed in this article outperforms the baselines in general. The results are highly encouraging because they have shown to be able to forecast well when compared to other methodologies used in the literature today.

Despite the fact that the input dimension is very large, the algorithm has proved that it can learn from it without the need of any dimension reduction techniques such as feature selection. When compared to other machine learning models, it has shown significant improvements in terms of accuracy; nonetheless, we feel that variance could be reduced, resulting in a more dependable model.

When it comes to financial results, it's vital to note that it was able to keep all stocks positive, even if the results weren't always the best when compared to the baselines.

Another plus is that it had a high return ratio per operation, implying that it was more successful in detecting large deviations, which is critical when accounting for transaction costs and taxes.

Furthermore, when comparing the maximum losses, it is reasonable to conclude that the LSTM-based model poses fewer dangers than the other options.

In Future, Instead of merely buying and selling after a set period of time, we plan to test the model using a variety of more realistic trading techniques. Take into account the inherent characteristics of stock markets, including as time, execution booking, and transaction costs.

REFERENCES

- Stock Price Prediction Using LSTM on Indian Share Market ,Achyut Ghosh1, Soumik Bose1, Giridhar Maji2, Narayan C. Debnath3, Soumya Sen1.
- K. A. Althelaya, E.-S. M. El-Alfy and S. Mohammed, "Evaluation of bidirectional lstm for short-and long-term stock market prediction", 2018 9th international conference on information and communication systems (ICICS), pp. 151-156, 2018.
- Future Stock Price Prediction using LSTM Machine Learning Algorithm Mrs. Nivethitha1, Pavithra.V2, Poorneshwari. G3, Raharitha. R4.
- B.G. Malkiel, A Random Walk Down Wall Street, New York:Norton, 1973.
- Illustrated Guide to LSTM's and GRU's: A step by step explanation Available:https://towardsdatascience.com/illustrated-guide-to-lstmsand-gru-s-a-step-by-step-explanation-44e9eb85bf21
- A. V. Devadoss and T. A. A. Ligori, "Forecasting of stock prices using multi layer perceptron", Int J Comput Algorithm, vol. 2, pp. 440-449, 2013. (https://scholar.google.com/scholar?as_q=Forecasting+of+stock+prices+using+multi+layer+perceptron&as_occt=title&hl=en&as_sdt=0%2C31)
- Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009). "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition"
- K. Chen, Y. Zhou and F. Dai, "A lstm-based method for stock returns prediction: A case study of china stock market", Big Data (Big Data) 2015 IEEE International Conference on, pp. 2823-2824, Oct 2015.
- L. Khaidem, S. Saha and S.R. Dey, Predicting the direction of stock market prices using random forest, CoRR, vol. abs/1605.00003, 2016, [online] Available: http://arxiv.org/abs/1605.00003.
- A. Sharang and C. Rao, Using machine learning for medium frequency derivative portfolio trading, CoRR, vol. abs/1512.06228, 2015, [online] Available: http://arxiv.org/abs/1512.06228.

- V. K. Menon, N. C. Vasireddy, S. A. Jami, V. T. N. Pedamallu, V. Sureshkumar and K. Soman, "Bulk price forecasting using spark over nse data set", International Conference on Data Mining and Big Data, pp. 137-146, 2016.
- S. A. Verma, G. Thampi and M. Rao, "Inter-comparison of artificial neural network algorithms for time series forecasting: Predicting indian financial markets", International Journal of Computer
- Applications, vol. 162, no. 2, 2017.
- M. S. Babu, N. Geethanjali and B. Satyanarayana, "Clustering approach to stock market prediction", International Journal of Advanced Networking and Applications, vol. 3, no. 4, pp. 1281, 2012.
- M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques", 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), pp. 322-327, 2016.
- R. Akita, A. Yoshihara, T. Matsubara and K. Uehara, "Deep learning for stock prediction using numerical and textual information", 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1-6, 2016.
- S. Selvin, R. Vinayakumar, E. Gopalakrishnan, V. K. Menon and K. Soman, "Stock price prediction using lstm rnn and cnn-sliding window model", 2017 international conference on advances in computing communications and informatics (icacci), pp. 1643-1647, 2017.
- P. Abinaya, V. S. Kumar, P. Balasubramanian and V. K. Menon, "Measuring stock price and trading volume causality among nifty50 stocks: The toda yamamoto method", Advances in Computing Communications and Informatics (ICACCI) 2016 International Conference, pp. 1886-1890, 2016.