Stock Market Forecasting

Zeel Prajapati[1], Varad Bhogayata[2], Shrutwa Shah[3], Prerak Shah[4], Prerak Raja[5], Fenny Zalavadia[6]
Arizona State University, Tempe, AZ
[1]zprajap1@asu.edu, [2]vbhogaya@asu.edu, [3]sshah120@asu.edu, [4]pshah70@asu.edu, [5]pkraja@asu.edu, [6]fzalavad@asu.edu

1. INTRODUCTION

Stock market is a platform where the buying and selling of shares takes place, mainly for the publicly listed companies. The stock market has a huge impact on the world economy, and it works as a barometer for the economy. Moreover, stocks are the best known form of financial tool for creating wealth and hence have been gaining the attention of even the mass public since the last few decades.

Stock Market Prediction (SMP) helps determine the future value of a company stock and other financial instruments traded on an exchange.[1] The investors are always looking to invest in companies with high growth potential, and hence the whole idea of predicting stock prices is to gain significant profits. Considering the ever changing market trends which do not only depend on the historical trends but also on the company's current standing, societal view of the company, social change, etc., makes it crucial to understand what changes the market and to what extent as well as to have a precise and accurate prediction of the future value of the stocks and the future trend.

With the recent advances in technology in the field of artificial intelligence and machine learning, several attempts have been made to predict the stock market and provide the future trend on the basis of previous historical data.[2] Machine learning has been a valuable tool to the domain of stock prediction due to its efficient and accurate measurements. The research in [3] uses the Support Vector Machine (SVM) for the prediction of stock prices of the upcoming day. Several fuzzy models have also been developed for SMP as in [4]. However, all these techniques were implemented considering only a few parameters, whereas there are quite a few more factors on which the stock market prices depend.

This project was implemented in an effort to understand the detailed implementation and accuracy of the machine learning algorithms for SMP using not just historical data but also other factors which make the share prices dynamic and volatile. Along with the machine learning algorithms, the project here uses several ensemble deep learning techniques as well, for making the predictions. Upon comparison between the two, deep learning techniques turned out to be the clear winners. The system also provides a dashboard which makes it easier for the investors and the mass public to visualize the changes and the future trend of the stock market easily.

2. PROBLEM DESCRIPTION

A rise or a fall in the prices of shares indicates what cycle the economy is in - recession or boom. Furthermore, the stock market provides a platform for an individual investor to invest his/her income and earn a share of the company's profit and in turn poses as a way of generating personal wealth and increasing investment. The investment market has seen a huge increase in the number of investors, who have to figure out the market themselves and make informed decisions on their own. This makes investment very stressful in modern societies. Most of the investors do not have access to sophisticated models or technologies which can help them make informed decisions.

Stock Market Prediction (SMP) is a topic in which the investors as well as the mass public takes great interest nowadays. The goal of SMP is to accurately predict the change in the stock prices of a company which are impelled by volatile factors such as societal change, news and other social media and hence makes it difficult to get accurate predictions only on the basis of historical data.[5]

The view of the mass public on the current standing of a company and other current affairs is clearly reflected by the social networking sites. Moreover, the financial news has a great impact on the fluctuations in the stock prices. The system here uses techniques like sentiment analysis performed on the financial news data obtained by scraping news websites. For improved accuracy and performance of the prediction models, feature engineering is performed on the datasets. In the end the system here considers, the public sentiment, news and historical data like the closing price for performing SMP, to understand the market standing and its effect on the stock prices.

Implementing machine learning algorithms and ensemble deep learning techniques on the financial news data and historical data here, have posed to be a great tool for SMP. Our system makes use of these techniques to find hidden patterns in the stock prices and shows the comparison of results obtained by implementing 1) ML algorithms - Linear Regression, K - Nearest Neighbour and SVR and 2) Deep Learning models - LSTM, LSTM + GRU and Prophet, and these results validate the success of the proposed methodology. This will in turn serve as a tool for investors to see the stock market performance and will also give them a different perspective to make informed decisions about which stocks to invest in.

3. METHODOLOGY

The entire project was divided into multiple phases which are clearly illustrated in Figure 1. The tools and technologies used in each of the phases are explained in detail after the flowchart.

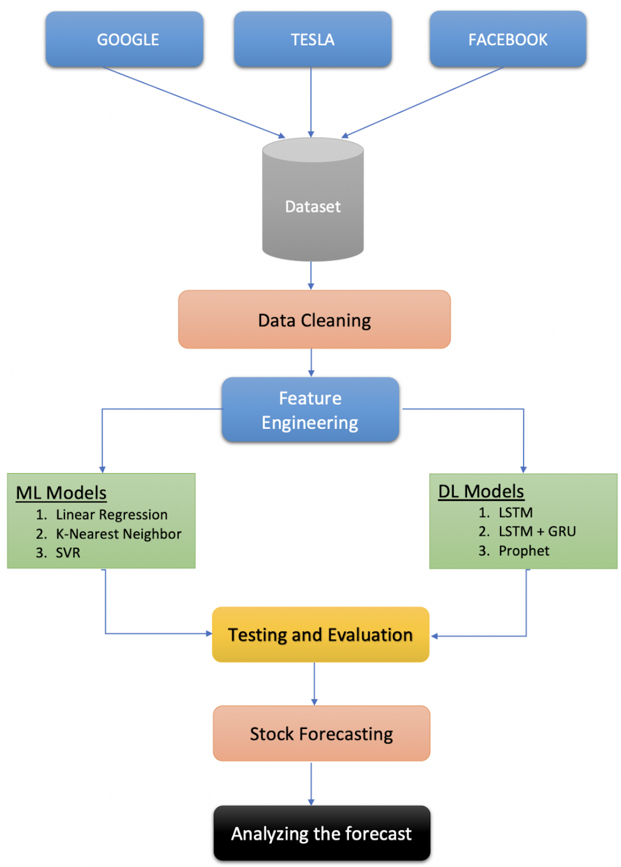


Figure 1. Methodology

3.1. Dataset

3.1.1 Stock Dataset

Stock data was collected from two sources: Yahoo Finance and Alpha Vantage API. Through the help of the get_intraday API provided by Alpha Vantage, time series data of the stock was collected. The intraday data is derived from the Securities Information Processor (SIP) market-aggregated data. The function get_intraday had the following parameters:

- symbol - NASDAQ code of the company
- interval - Time interval between two consecutive data points in the time series. The following values are supported: 1min, 5min, 15min, 30min, 60min

- outputSize - Strings compact and full are accepted with the following specifications: compact returns only the latest 100 data points in the intraday time series; full returns the full-length intraday time series.
- datatype - json or csv

Historical data about a stock was fetched from Yahoo Finance. The data was directly downloaded in .csv files and consisted of columns like Date, Open Price, Close Price, High Price, Low Price, Volume.

3.1.2 News Headlines Dataset

Required company's news headlines were extracted from https://finviz.com/. By using a web scraping library beautifulsoup news headlines were extracted.

The main reason behind including this new feature column is to consider features such as the company's current market standing, societal view of the company apart from historical data.

3.2. Preprocessing

3.2.1 Data Cleaning

Variable type of Date column was converted from string to timestamp.

Additionally, Close Price values were normalized using minmaxscaler.

3.2.2 Feature Engineering

Out of all price columns: Open, Close, High and Low, close price of stocks was of fundamental importance as it would be the final price for the stock for that time interval.

As it is not efficient to give the news headlines directly as an input to the model, their sentiment scores were fetched by using the vader API on top 100 headlines related to the company. Vader API gave output in four columns: positive sentiment score, negative sentiment score, neutral sentiment score and complex sentiment score. Since overall sentiment of news headlines was required, complex(combined) score was required.

3.3. Model

The key idea was to explore different machine learning and deep learning techniques and choose the best model for our stock price prediction.

3.3.1 Machine Learning

For this project, we have used three Machine Learning models namely, Linear Regression, K-Nearest Neighbour and Support Vector Regressor. Mainly, this was done using libraries like scikit-learn and pandas for preprocessing, training and testing. The fetched data is preprocessed by using sklearn's preprocessing library. From the library we have used the tool minmaxscaler to perform normalization on the input data. After that we have split the input data into train, validation and test sets. Then, through the help of the sklearn library, all the three models are instantiated and then trained on the available data.

3.3.2 Deep Learning

After exploring different Deep learning algorithms it was found that Long Short Term Memory (LSTM) performs better with time-series data. Hence the model was trained using LSTM. Here, our LSTM model has 4 LSTM layers and 1 dense layer. Also dropout regularisation in between the main layers was used.

Then, in order to incorporate the new sentiment score the Gated Recurrent Unit (GRU) model was used. An ensemble model was made using the LSTM and GRU model. Then, this ensemble model was trained on the input data.

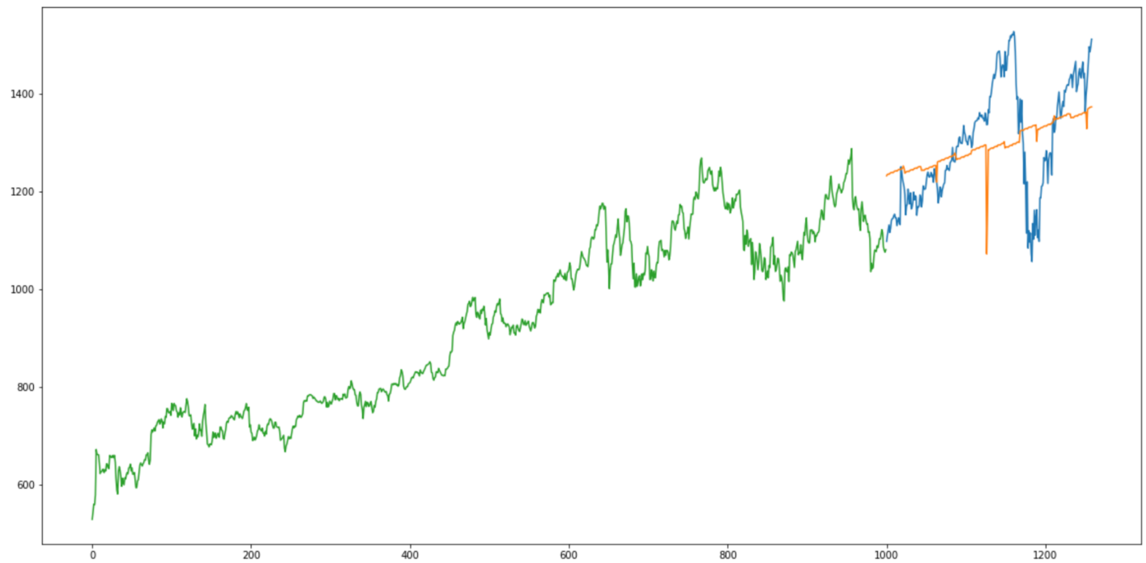
Additionally, a standard time series model provided by Facebook was also used. Through the help of the library “Prophet” the model was run on the input data and its results were compared with our machine learning and deep learning models.

4. RESULTS

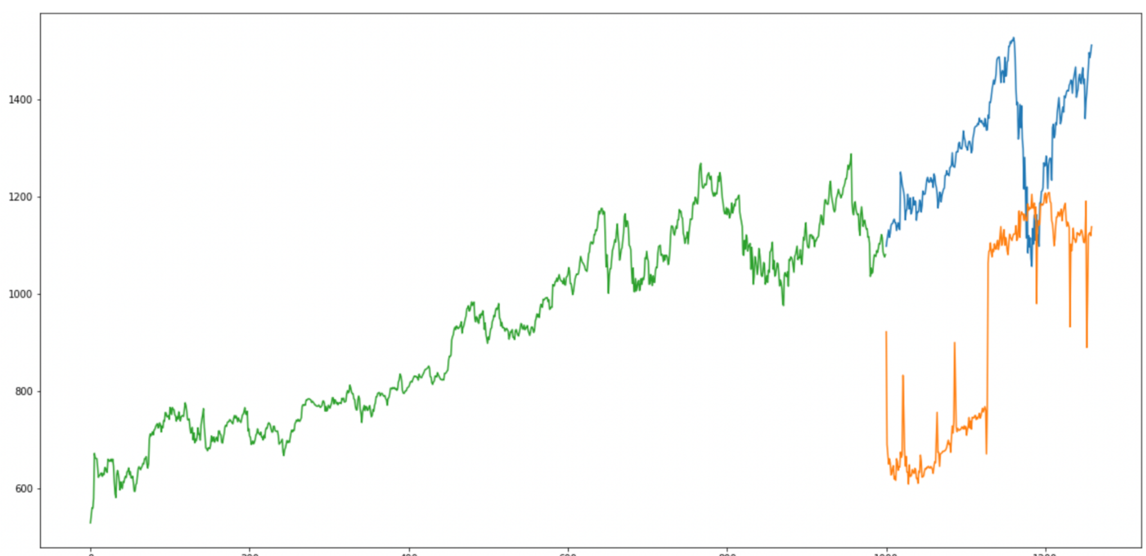
In order to evaluate the Machine Learning model and Deep Learning models, Root Mean squared error (RMSE) was used as an evaluation metric.

4.1 Machine Learning Models

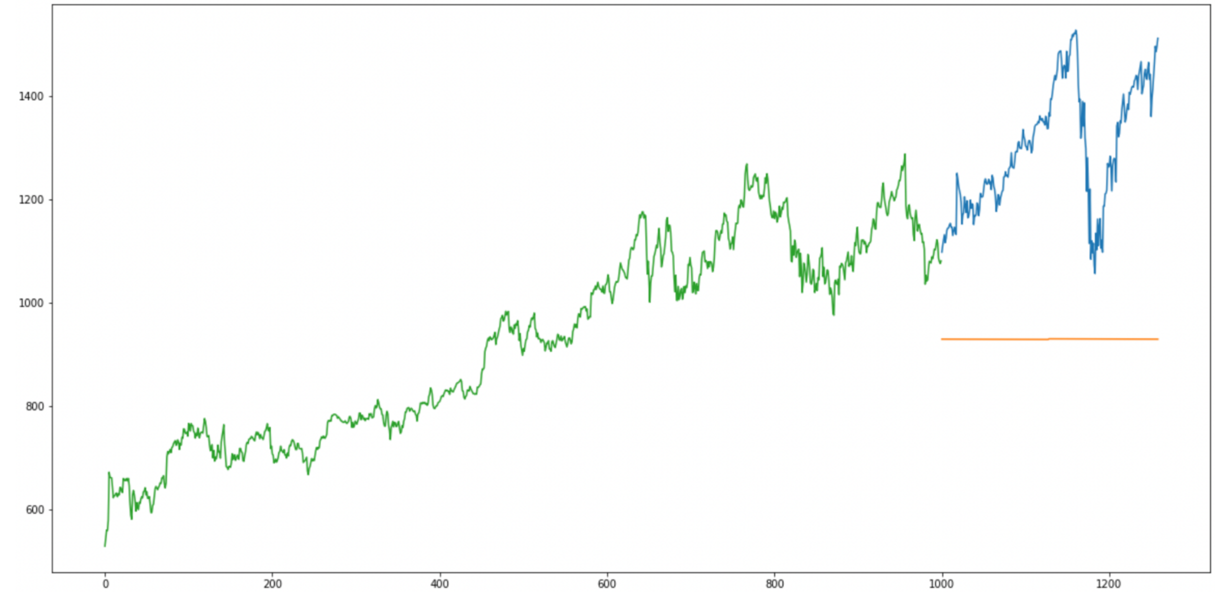
The following are the results of the Machine Learning models.



Linear Regression Model



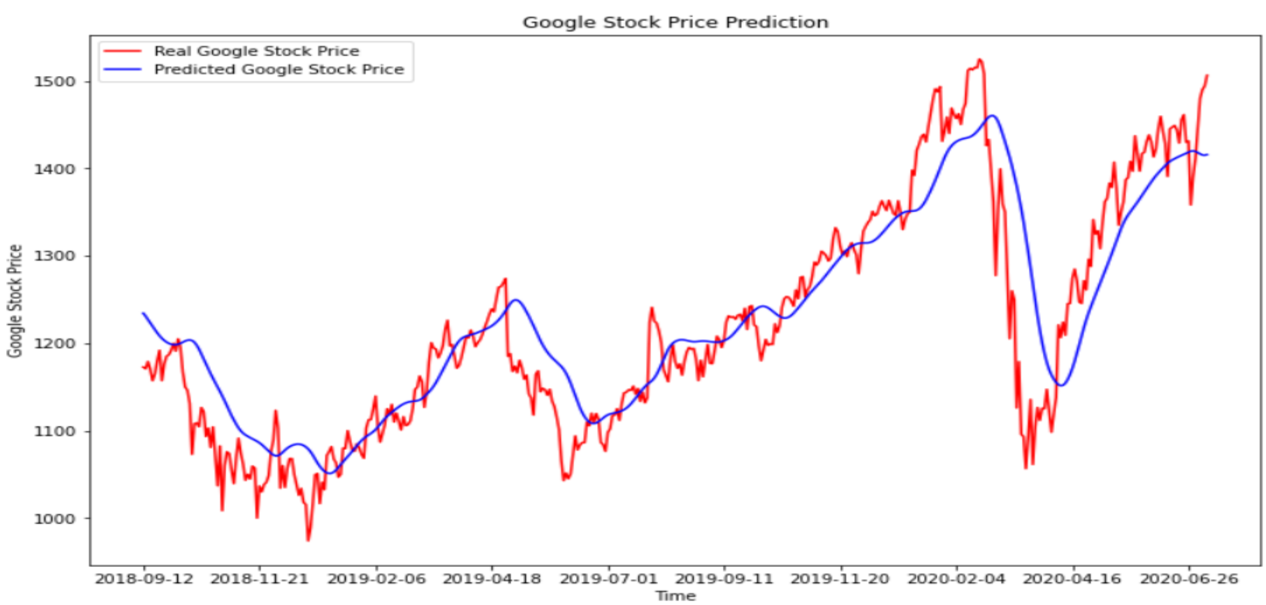
K-Nearest Neighbors Model



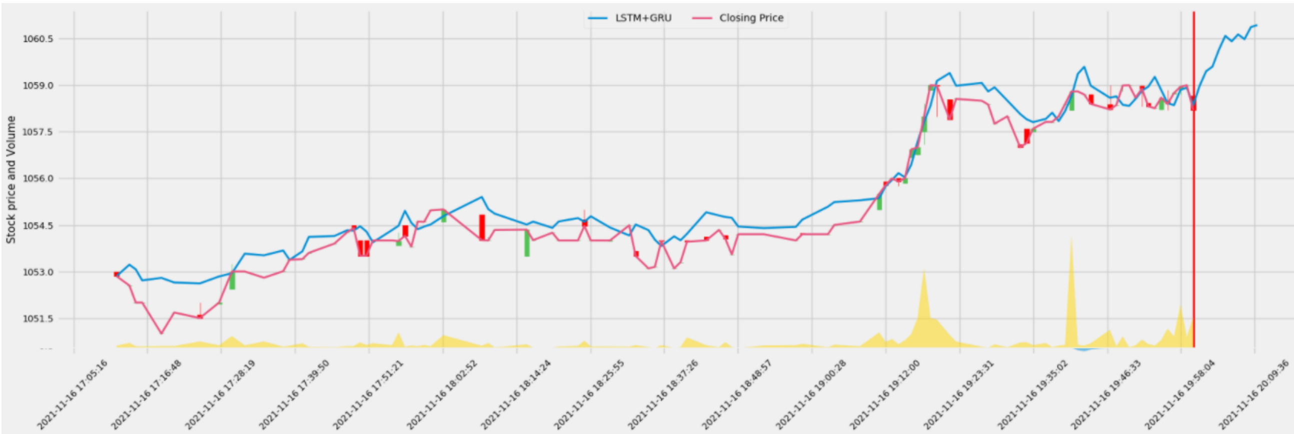
Support Vector Regression Model

In all of these 3 models the green line represents the historical stock price. After that the blue line represents the actual stock price and the orange line represents the stock price that the model predicts. By comparing the Root Mean Square Error (RMSE) score, it was found that from the three models it can be clearly seen that the linear regression model performs the best. However, it can clearly be seen from the Linear Regression Model graph that the model does not perform up to the mark.

4.2 Deep Learning Models



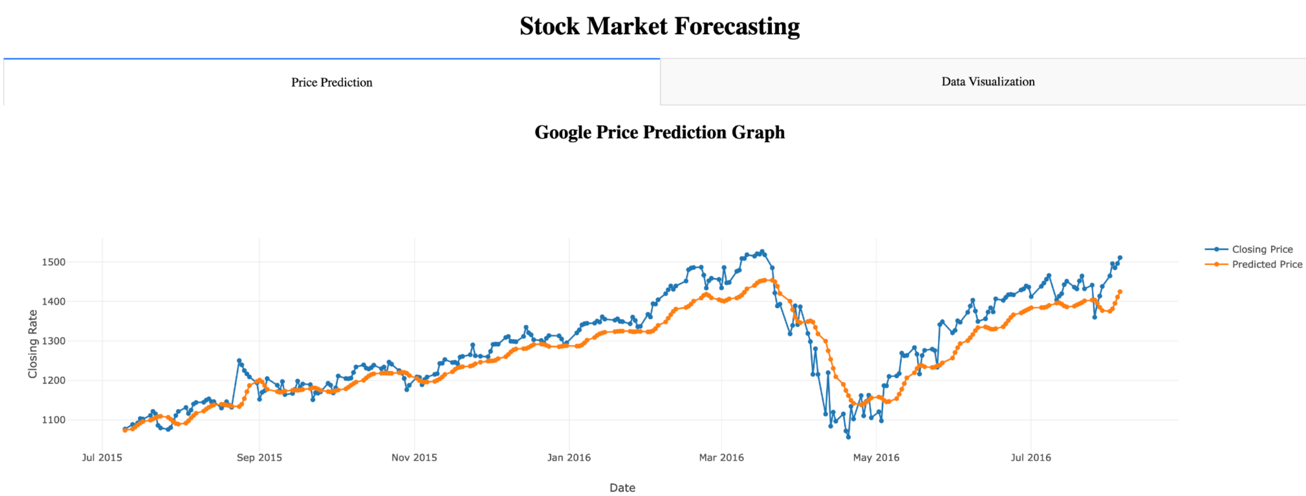
LSTM Model



LSTM + GRU Model

The above graph shows the result of our ensemble model (LSTM + GRU). The red one is the actual stock price. And the blue is predicted. The red line is the data that was available till then and after that the blue line is the future prediction.

For visualization and plotting the results, we have used matplotlib and plotly-dash libraries which are shown below.



Visualization Tool

5. CONCLUSION

In this project, we have tried to do the comparative studies between different Machine learning models and Deep learning models. We can conclude that Deep learning models like Long Short-Term Memory (LSTM), LSTM with Gated Recurrent Unit (GRU) outperforms the Machine Learning Models such as Linear regression, Support vector regression and KNN (K-Nearest Neighbor) by predicting the future stocks as accurately as the real value. Our system predicts the future stocks based on the most recent stocks, i.e predicted stocks for today will be highly biased on yesterday's value. As a result of which our model overfits and may perform poorly if we try to predict the stocks in future for a longer period. Along with opening and closing values of the stocks, we have tried to make use of sentiments from the company's news. In doing so we faced few complications while implementing it, like we had to explicitly extract the sentiments from news and generate scores out of it and add that as an extra feature to our model. Overall it was complex but as a result our model's accuracy was increased.

6. FUTURE WORK

Few improvisations we can do in the future is that along with news sentiments of the company, we can include twitter sentiments as a feature. Not only this we can further consider earnings, cash flow and income statements of the company to better predict the future stocks. Despite such results, the model was not able to work well with high volatility and hence value accuracy. More parameters can be considered or more models can be ensembled for more accurate results. For the dashboard created, it can be further scaled to include more companies' stocks, different types of visualizations can be added along with individual line plots of stocks, we can include bar plots which compare stocks of different companies and give users better insights. And lastly, with recent research and findings of new Machine and Deep learning models, we can build a Real-time stock market prediction system.

REFERENCES

[1] I. Parmar et al., "Stock Market Prediction Using Machine Learning," *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2018, pp. 574-576, doi: 10.1109/ICSCCC.2018.8703332.

[2] Henrique, Bruno Miranda, Vinicius Amorim Sobreiro, and Herbert Kimura. "Literature review: Machine learning techniques applied to financial market prediction." *Expert Systems with Applications* 124 (2019): 226-251.

[3] Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms." *Department of Electrical Engineering, Stanford University, Stanford, CA* (2012): 1-5.

[4] A. Sheta, "Software Effort Estimation and Stock Market Prediction Using Takagi-Sugeno Fuzzy Models," *2006 IEEE International Conference on Fuzzy Systems,* 2006, pp. 171-178, doi: 10.1109/FUZZY.2006.1681711.

[5] Khan, W., Ghazanfar, M.A., Azam, M.A. *et al.* Stock market prediction using machine learning classifiers and social media, news. *J Ambient Intell Human Comput* (2020). https://doi.org/10.1007/s12652-020-01839-w

[6] 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)*, 2016, pp. 322-327, doi: 10.1109/ICCOINS.2016.7783235.

[7] S. M. Idrees, M. A. Alam and P. Agarwal, "A Prediction Approach for Stock Market Volatility Based on Time Series Data," in *IEEE Access*, vol. 7, pp. 17287-17298, 2019, doi: 10.1109/ACCESS.2019.2895252.

[8] A. Shakva, A. Pokhrel, A. Bhattarai, P. Sitikhu and S. Shakva, "Real-Time Stock Prediction Using Neural Network," *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2018, pp. 1-4, doi: 10.1109/CONFLUENCE.2018.8443057.