Comparative Analysis of Apple Stock Price Prediction Methods

Vivek Haridas
AM.EN.U4AIE19071
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
vivekharidas@am.students.amrita.edu

***Abstract*— Investors in the stock market have always been on the lookout for fresh and distinctive approaches to correctly anticipate stock price movement and make a large profit in recent years. However, investors continue to hunt for improved and innovative strategies to beat the market rather than old and traditional ones. As a result, experts are always attempting to develop fresh approaches to meet investor demand. In this project, we will implement Stock Price Prediction along with a comparative study of multiple models such as RNN, ARIMA, MA, and GAN, and examine their efficiency using various evaluation metrics such as RMSE, MAPE, MAE, and MPE metrics.**

Index Terms - Forecast, ARIMA, GAN, LSTM, BiLSTM, GRU, GAN, MA, XGBoost, MAE, RMSE, MPE, MAPE.

I. INTRODUCTION

Stock market prediction is the process of trying to anticipate the future value of a company's stock that is traded on an exchange. Accurately forecasting the stock market is akin to being able to see into the future. Stock is a financial commodity that many people like because of its high risk, great reward, and flexibility in trading. Investors may make a lot of money by correctly predicting stock price patterns. However, the stock price is impacted by a variety of factors, including the macroeconomic state, market conditions, key social and economic events, investor preferences, and managerial actions made by corporations. As a result, stock price prediction has long been a priority and demanding study area. Traditional stock price prediction methods rely on statistical and econometric models, but these models are incapable of dealing with the stock market's dynamic and complicated environment. The observation of changes in the trend of data over time is referred to as time series analysis. Time series analysis has several uses. One such use is the prediction of an item's future value based on historical values. The finest example of such an application is definitely future stock price prediction. Manually anticipating stock price patterns based on stock data is a time-consuming operation.The stock price is impacted by a variety of unpredictability variables, resulting in typical non-stationary stock price time-series data. As a result, stock price prediction is one of the most difficult issues in all of prediction research. Scholars have researched stock price prediction from a variety of approaches throughout the last several decades. The two most important directions are the enhancement of prediction models and the selection of model characteristics. However, with the introduction of artificial intelligence, it is now feasible to anticipate the stock market using massive data and better computational skills.

In this paper, we will look at how to do time series analysis using several models. We will forecast the future stock values of the Apple Company (AAPL) using daily and weekly data from January 1st, 2014 to April 1st, 2022.

II. RELATED WORKS

The study by Kavinnilaa J, Hemalatha E, Minu Susan Jacob and Dhanalakshmi R on "Stock Price Prediction Based on LSTM Deep Learning Model" asserts that historical value takes into account all other market events can be used to forecast future movement. Machine Learning techniques can identify methodologies and perspectives that can be used to make surprisingly accurate predictions. The LSTM model is chosen to analyse and forecast stock market prices in order to make more informed and precise financial decisions. The proposed method utilizes a close price from S&P 500 component shares, to estimate the future close price.

The paper by Soheila Mehrmolaei and Mohammad Reza Keyvanpour on "Time series forecasting using improved ARIMA" lists past important research that examined time series data forecasting in many application areas. They suggest an unique method for improving the ARIMA model for time series forecasting by using a mean of estimate error. The experimental findings show that the suggested technique can increase performance in the time series data forecasting process. Based on predicting length, time series forecasting approaches have been categorised and divided into two classes. Short-term and long-term estimators have been named by several groups. Furthermore, for time series forecasting in the basic ARIMA model, a strategy based on applying a mean of estimation error is presented. As a consequence, they can state that the modified ARlMA model is superior than the basic ARIMA model, and the suggested technique has acceptable accuracy.In their work, three time series data sets are utilised to evaluate the performance of the suggested technique to basic ARIMA. The assessment results show that the proposed technique is adequate and successful for uni variate time series forecasting.

In the paper written by Jie Jiang about "Stock Market Prediction Based on SF-GAN Network" he presents SF-GAN, a new Generative Adversarial Networks (GAN) architecture for forecasting stock closing prices and trends. The State Frequency Memory Neural Network (SFM) is used as the generator, while the Convolutional Neural Network (CNN) is used as the discriminator, in SF-GAN.The SFM is a State Frequency Memory Neural that can dynamically learn state frequencies.The proposed architecture performs better than state of the art models such as LSTM based on their MAE score.The experimental findings suggest that the SF-GAN model developed in this work may increase stock price trend prediction accuracy and minimise price prediction error.

III. DATASET

The project uses the APPLE stock price data from the Yahoo Finance API as the dataset. In order to analyse the prediction by different models from different perspectives, we will be considering three sets datasets:

1) Daily Stock Price Dataset with timeline ranging from 1st January 2014 to 4th April 2022.
2) Weekly Stock Price Dataset with timeline ranging from 1st January 2014 to 4th April 2022.
3) Stock Price Dataset with data from January 2016 to August 2019 along with scraped tweets and their sentiment polarity to be used as a feature.

To create the third dataset, we scrape for tweets using the tweepy API and do sentiment analysis on the tweets to determine the tweet polarity. This is then used as a feature for prediction.

Using these three sets of data, we compare and contrast the performance of different models.

IV. METHODOLOGY

Stock price prediction is a difficult job that is simulated using machine learning to anticipate stock returns. There are several approaches and instruments available for stock market forecasting. The stock market is thought to be extremely active and complicated. A more accurate projection of future prices might result in a better profit return for stock investors. We will be doing a comparative analysis using the models like LSTM, Bi-directional LSTM, GRU, ARIMA, simple moving average and exponential weighted moving average, GAN and so on to determine the best model, based on the error accuracy measures like RMSE, MAE, MAPE, and MPE .

A. RNN

Recurrent Neural Network (RNN) is an advanced form of neural networks that has internal memory which makes RNN capable of processing long sequences. This makes RNN very suitable for stock price prediction, which involves long historical data. We have used variants of RNN models including LSTM, BLSTM, and GRU models.

1) Long short-term memory(LSTM): Long short-term memory (LSTM) is a deep learning artificial recurrent neural network (RNN) architecture. Unlike traditional feedforward neural networks, LSTM has feedback connections. It is capable of processing not just single data points, but also complete data sequences. A simple LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. The cell remembers values across arbitrary time periods, and the three gates control the flow of information into and out of the cell.

In our LSTM model, we have two LSTM layers with 64 units each followed by two Dense layers. We use Adam for optimisation and mean squared error as the loss function.

2) Bi-LSTM: Bidircectional LSTM is a model in which, the neural network takes the sequence information in both directions (forward and backward). The input is fed bidirectionally and flows in two ways, which separates a bidirectional LSTM from a conventional LSTM. For this, the BiLSTM architecture has trains two models rather than just one. The first model is trained to learn the sequence of the provided input, while the second model learns the sequence of input in the reverse direction. As there are two trained models, a mechanism called the merge step was created to combine them. Merging can be done using methods like concatenation (default method), addition, multiplication, averaging etc

In our BiLSTM model, we have one Bidirectional LSTM layer and one LSTM layer each of which has 64 units followed by two Dense layers. We use Adam for optimisation and mean squared error as the loss function.

3) GRU: Gated Recurrent Unit (or GRU in short) is a recurrent neural network that seeks to tackle the vanishing gradient problem. Because both are constructed similarly and, in certain situations, achieve equally outstanding outcomes, GRU might be regarded a variant of the LSTM. GRU employs the so-called update gate and reset gate to overcome the vanishing gradient problem of a regular RNN. In essence, there are two vectors that determine what data should be sent to the output. They may be taught to retain knowledge from the past without having it washed away over time, and to discard information that is unnecessary to the forecast.

In our GRU model, we have one GRU layer and one LSTM layer each of which has 64 units followed by two Dense layers. We use Adam for optimisation and mean squared error as the loss function.

B. ARIMA

ARIMA stands for autoregressive integrated moving average. It is a statistical analysis model that uses time series data to better understand the data set or anticipate future trends. If a statistical model predicts future values based on previous values, it is called autoregressive. For example, an ARIMA model may try to estimate a company's earnings based on prior periods or predict a stock's future pricing based on historical performance. The purpose of the model is to anticipate future securities or financial market movements by looking at the discrepancies between values in a series rather than actual values. Autoreggresion(AR) is a model in which a changing variable regresses on its own lagged, or previous values. The term "integrated" refers to the differencing of raw observations in order for the time series to become stationary. Moving average (MA) includes the reliance between an observation and a residual error from a moving average model applied to lagged data.

The ARIMA model is defined by parameters (p,d,q), where p represents the number of autoregressive terms, d represents the number of nonseasonal differences required for stationarity, and q represents the number of lagged forecast errors in the forecast equation. In our model, we used parameters of (4,1,0).

C. MA

A moving average (MA) is computed by combining current prices by the number of time periods in the computation average. The average is calculated over a set length of time, such as 10 days, 20 minutes, 30 weeks, or any other time period selected by the trader. There are benefits to employing a moving average in your trading, as well as many types to choose from. Moving average methods are very popular and can be customised to any time period, making them suitable for both long and short-term investors and traders. There are several techniques to compute a moving average. A five-day simple moving average (SMA) calculates a new average each day by adding the five most recent daily closing prices and dividing the total by five. The single flowing line is created by connecting each average to the next. The exponential weighted moving average is another common form of moving average (EWMA). The formula is more complicated since the most recent prices are given greater weight.

For our project, we have implemented and compared SMA models with window sizes 10 and 20 along with EWMA models with spans of 3, 6 and 12.

D. GAN

GANs, or Generative Adversarial Networks, are a type of generative modelling that use deep learning techniques such as convolutional neural networks. In machine learning, generative modelling is an unsupervised learning job that entails automatically detecting and learning regularities or patterns in input data such that the model may be used to produce or output new instances that might have been drawn from the original dataset.A generator and a discriminator are both present in GANs. The Generator creates fictitious data samples in order to deceive the Discriminator. On the other hand, the Discriminator seeks to discriminate between actual and fraudulent samples. Both the Generator and the Discriminator are Neural Networks, and throughout the training phase, they compete with each other. The procedures are performed multiple times, and with each iteration, the Generator and Discriminator improve their performance in their respective roles.

E. XGBoost

XGBoost is an ensemble Machine Learning technique based on decision trees that employs a gradient boosting structure. Artificial neural networks outperform all other algorithms or frameworks in prediction problems concerning unstructured data. However, decision tree-based algorithms are currently considered best in class for small to moderate structured or tabular data. Gradient boosting is an iterative process for converting weak classifiers to strong learners. The name XGBoost alludes to the engineering goal of pushing the computational resource limit for boosted tree methodologies.

We use this model to compare the effect of tweet sentiment on stock price.

V. RESULTS

For comparing the performance of models, we used predictive accuracy measures like mean absolute error (MAE), root mean squared error (RMSE), mean percentage error (MPE) and mean absolute percentage error (MAPE).

We begin by evaluating the models implemented on the Daily stock price dataset.

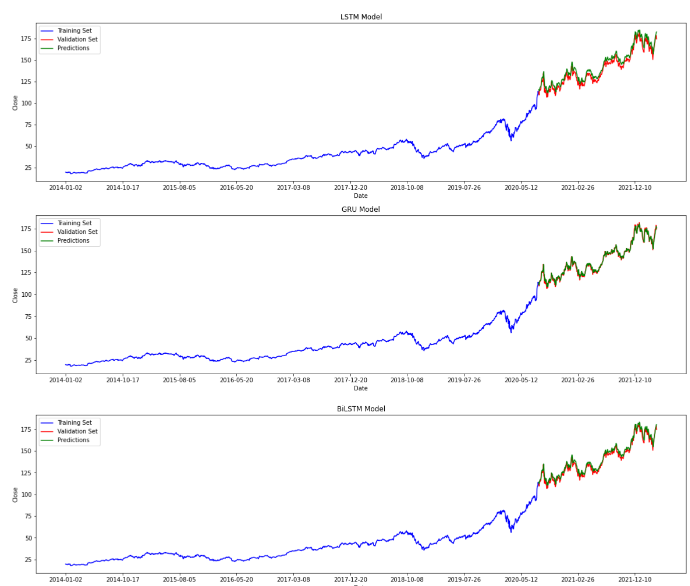

Fig. 1. RNN Models on Daily Dataset

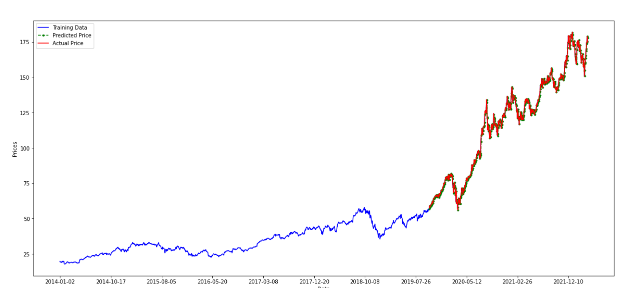

Fig. 2. ARIMA on Daily Dataset

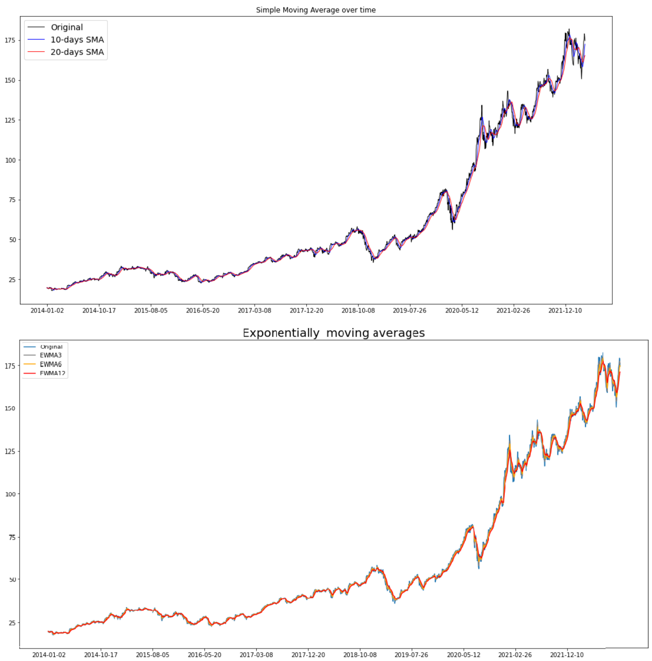

Fig. 3. SMA and EWMA on Daily Dataset

Model	MAE	RMSE	MPE	MAPE
ARIMA	1.76	2.43	-0.17%	1.55%
SMA-20	2.12	3.33	0.85%	3.35%
EWMA-12	1.33	2.14	0.49%	2.10%
LSTM	4.43	5.08	3.09%	3.22%
GRU	2.02	2.65	0.38%	1.47%
BiLSTM	2.98	3.63	1.82%	2.17%
GAN	1.29	1.84	0.03%	2.28%

Next, we evaluate the models implemented on the Weekly stock price dataset.

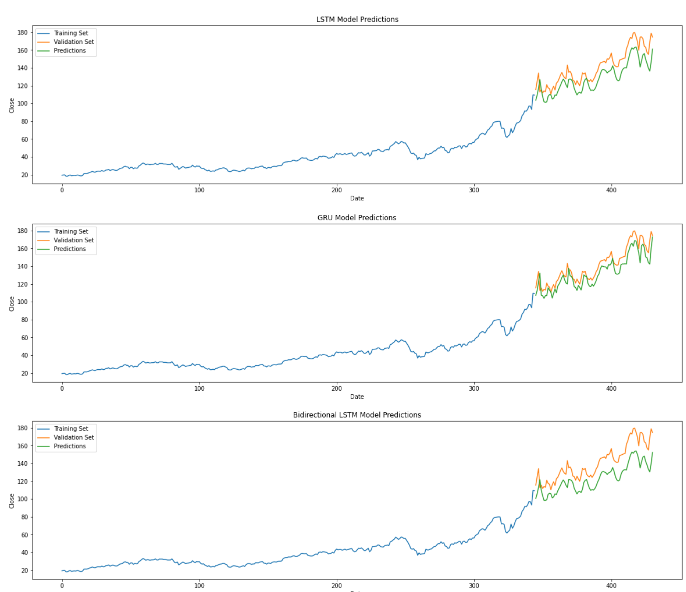

Fig. 4. RNN Models on Weekly Dataset

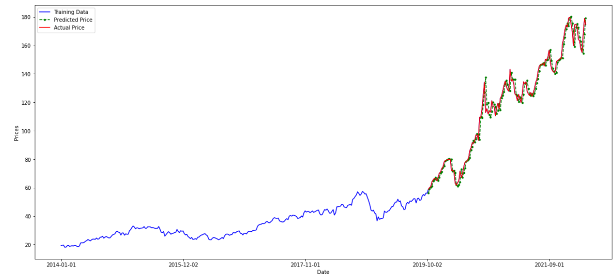

Fig. 5. ARIMA on Weekly Dataset

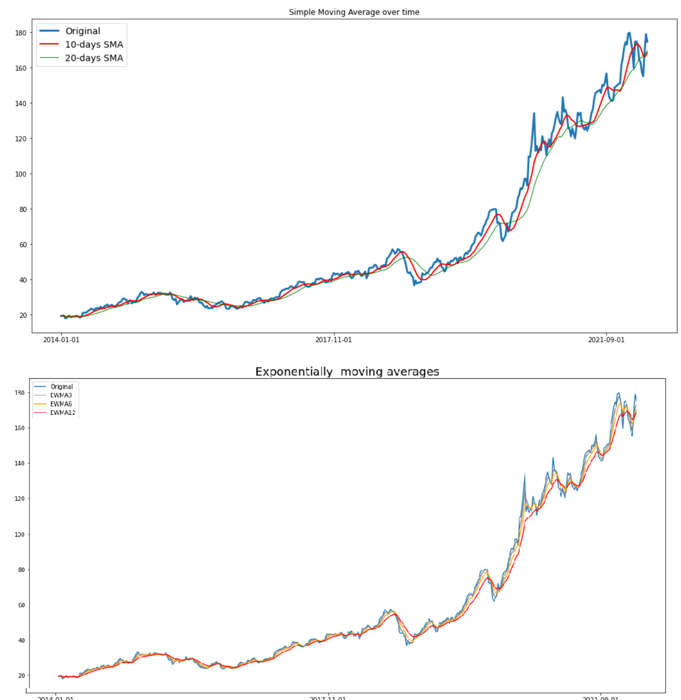

Fig. 6. SMA and EWMA on Weekly Dataset

Model	MAE	RMSE	MPE	MAPE
ARIMA	3.82	5.41	-0.54%	3.25%
SMA-20	4.80	7.38	4.06%	7.62%
EWMA-12	3.15	4.78	2.34%	4.95%
LSTM	12.45	14.04	-8.38%	8.68%
GRU	8.90	10.44	-5.79%	6.27%
BiLSTM	19.07	19.37	-12.41%	12.59%
GAN	1.44	2.09	4.64%	7.51%

Finally we look at the models implemented using the sentiment polarity of the tweet as a feature and compare with the same model when not using the sentiment polarity of the tweet.

Model	MAE	RMSE	MPE	MAPE
XGBoost (with Sentiment)	0.26	0.33	0.42%	0.77%
XGBoost (without Sentiment)	0.36	0.43	0.51%	0.83%

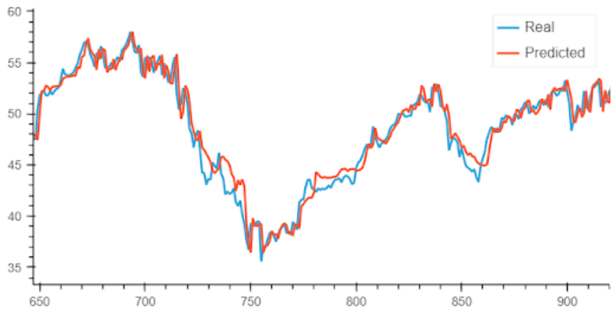

Fig. 7. XGBoost Model using Polarity as feature

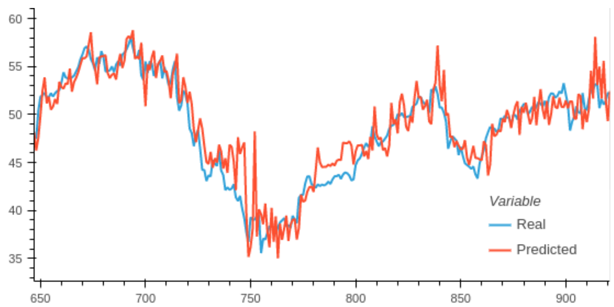

Fig. 8. XGBoost Model without using Polarity as feature

VI. CONCLUSION

From the above results, we can conclude that ARIMA and GAN were the best models for forecasting the stock prices for both Daily and Weekly Data. In the case of the dataset using sentiment polarity of tweets, we saw that the model that used the tweet polarity had predicted stock prices better than the same model that did not use tweet polarity as a feature.

VII. FUTURE SCOPE

In the future, we plan to assign different weights to the polarity of tweets based on different factors like the relevance of tweeter, number of retweets and so on to not give equal importance to all tweets. We also plan to implement a system where the tweets made after a particular time frame (based on when the stock market closes during the day) will only be considered for forecasting the next day and also take into consideration the tweets made during weekends (when the stock market is closed) on the subsequent opening day.

REFERENCES

[1] Kavinnilaa J, Hemalatha E, Minu Susan Jacob, Dhanalakshmi R, “Stock Price Prediction Based on LSTM Deep Learning Model”, 30 July 2021.
[2] Soheila Mehrmolaei, Mohammad Reza Keyvanpour, “Time series forecasting using improved ARIMA”, 9 April 2016.
[3] Jie Jiang, “Stock Market Prediction Based on SF-GAN Network”, 11 January 2021.
[4] Yan, W. ZHong, H. Qiu and Y. Xue, "Gaussian process for long-termtime series forecasting", 14 June 2009.
[5] S. Mehrmolaei and M. R. Keyvanpour, "A Brief Survey on Event Prediction Methods in Time Series", January 2015.
[6] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," 2015.
[7] Yue Yang, Yang Wu, Peikun Wang and Xu Jiali, "Stock Price Prediction Based on XGBoost and LightGBM", 2021
[8] Shoban Dinesh, Nithin Rao R, S P Anusha, Samhitha R, "Prediction of Trends in Stock Market using Moving Averages and Machine Learning", 2 April 2021.