STOCK PRICE PREDICTION USING SENTIMENT ANALYSIS AND TIME SERIES

A DISSERTATION

SUBMITTED TO

DEPARTMENT OF STATISTICS

SVKM’S NMIMS (DEEMED -TO-BE UNIVERSITY)

IN PARTIAL FULFILMENT FOR THE DEGREE OF

**MASTER OF SCIENCE**

**IN STATISTICS AND DATA SCIENCE**

BY

SUKHADA SAKHALKAR

PRANJAL GAHANKARI

MALAY SAHU

KEYA DOSHI

HRISHIKESH GURAV

UNDER THE SUPERVISION OF

**DR. LEENA KULKARNI**

**SUNANDAN DIVATIA SCHOOL OF SCIENCE**

SVKM’S Narsee Monjee Institute of Management Studies

(Deemed-To-Be University)

Vile Parle West, Mumbai, Maharashtra 400056

NOVEMBER, 2022

**ACKNOWLEDGEMENT**

We would like to express our warmest thanks of gratitude to our supervisor Dr. Leena Kulkarni for providing us the opportunity to do this project on Stock Price Prediction using Sentiment Analysis and Time Series. Their invaluable guidance and advice helped us to carry out the project successfully.

Also, we would like to thank our course coordinator Dr. Pradnya Khandeparkar for providing necessary facilities required for completion of this project.

We would also like to extend our gratitude to the Department of Statistics for guiding and supporting us at different stages of our project.

Finally, with great support of our team members and excellent guidance of mentors we are able to complete our project within a specific time.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Topic** | **Page No** |
| 1 | Introduction | 4 |
| 2 | Need for Study | 4 |
| 3 | Abstract | 5 |
| 4 | Literature Review | 5 |
| 5 | Objectives | 6 |
| 6 | Problem statement | 6 |
| 7 | Data Collection | 6 |
| 8 | Data Description | 7 |
| 9 | Research Methodology | 9 |
| 10 | Data Pre-processing | 10 |
| 11 | Techniques Used | 11 |
| 12 | Model Findings | 14 |
| 13 | Conclusion | 30 |
| 14 | Future Scope | 30 |
| 15 | References | 31 |
| 16 | Glossary and Terminologies | 31 |

**INTRODUCTION**

Forecasting any financial asset is a charming and challenging problem at times. The stock prices are affected by various factors. Some of the important factors can be considered as tweets by influential persons, historical prices. Consequently, the purpose of this work is to apply appropriate analytical techniques to them in order to make better stock price predictions.

The Indian stock market has always drawn attention and served as a source of investment for almost everyone. The Stock Market positions itself as a challenge to draw people in and encourage them to invest their money and income in stocks, whether they succeed or fail. The main obstacle the market presents to investors is its volatility and spontaneity. The market experiences a variety of emotions, including enthusiasm, joy, craftiness, regret, and many others, which it communicates through its investors.

Sentiment is defined as an attitude or opinion that is often caused or influenced by emotion. Sentiment analysis is the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Recurrent neural networks, such as LSTM, are frequently employed to solve problems involving sequence prediction. It has the capacity to remember past information that is significant while erasing unimportant information. This machine learning approach is unsupervised. As for how this model will work, it will train on historical stock price data, analyse that data, and forecast future stock values. Research work helps investors to make decisions based on the analytical evidence provided.

### NEED FOR THE STUDY

The stock market fluctuates daily, so profiting from the stock market requires careful planning. Since the inception of the stock market, analysts have always had difficulty predicting the future price of shares due to their complexity and profitability. The easiest and most reliable way to predict the future is to try to understand the present, but the amount of data available these days is staggering.

Data analysis is used to better understand the current scenario of the stock market and try to understand and create better future ranges for stock investments. It can add some certainty to the nature of stable stock prices. This certainty goes a long way in minimising losses and maximising profits. The forecast will never be perfectly accurate, but even a slight improvement in the accuracy of the forecast will increase the profitability of the lot.

Motivation to use methods of LSTM, RNN, sentiment analysis to determine if there are long-term dependencies on the data present in a given dataset. This can be determined from performance models. LSTM and RNN architectures for identifying long-term dependencies are used to predict the future.

**ABSTRACT**

Objective of this research is to build a model to predict stock price using sentimental information from tweets of twitter and historical prices, and the model is able to not only conclude better results but also minimise the difference between predicted values and actual values. Also, we checked if there is any impact of tweets on stock price. We apply various approaches to extract information from twitter. On the other hand, price data through time series are also useful to predict stock prices. Hence, improvement is made with a combination of sentiment analysis of twitter and time series analysis of historical prices.

**LITERATURE REVIEW**

Over the years, many papers have been published in the field of stock price prediction and sentiment analysis but not all of them look at a combined view of the both.

Objective of this research is to build a model to predict stock price using sentimental information from news headlines and historical prices, and the model is able to not only conclude better results but also minimise the difference between predicted values and actual values ([1]). This research is focused on proposing a hybrid deep learning (DL) based predictive model, that combines a Bidirectional Cuda Deep Neural Network Long Short-Term Memory (BiCuDNNLSTM) and a one-dimensional Convolutional Neural Network (CNN), for timely and efficient prediction of stock prices. Our proposed model (BiCuDNNLSTM-1dCNN) is compared with other hybrid DL-based models and state of the art models for verification using five stock price datasets ([2]).

This paper studies to what extent public Twitter sentiment can be used to predict price returns for the nine largest cryptocurrencies: Bitcoin, Ethereum, XRP, Bitcoin Cash, EOS, Litecoin, Cardano, Stellar and TRON. By using a cryptocurrency-specific lexicon-based sentiment analysis approach, financial data and bilateral Granger-causality testing, it was found that Twitter sentiment has predictive power for the returns of Bitcoin, Bitcoin Cash and Litecoin ([3]). We employ a Convolutional Neural Network model for classifying the investors’ hidden sentiments, which are extracted from a major stock forum. We then propose a hybrid research model by applying the Long Short-Term Memory (LSTM) Neural Network approach for analysing the technical indicators from the stock market and the sentiment analysis results from the first step ([4]).

This paper uses a recently introduced model for predicting stock price. This proposed model is a well-liked model named the Recurrent Neural Network (RNN) model. The major goal of this article is to determine to what degree a Machine Learning algorithm can anticipate the stock market price with greater accuracy ([5])In this paper, we apply sentiment analysis and machine learning principles to find the correlation between” public sentiment” and” market sentiment”. We use twitter data to predict public mood and use the predicted mood and previous days’ DJIA values to predict the stock market movements ([6]).

**OBJECTIVES**

* Stock price prediction using Time Series volatility model.
* Understanding the public’s sentimental view.
* Study the relationship between the historical prices and the sentiment score.
* Study the effect of tweets on the stock price prediction.

**PROBLEM STATEMENT**

“Buy LOW, sell HIGH “, this is the strategy used by most of the investors. But this is not the best strategy that should be used. Before investing investors should study how that particular stock is behaving. It could be very dangerous if investors invest in the good stock at bad times. Understanding which stock to buy and what time to sell that stock is a critical problem investors face.

**DATA COLLECTION**

The data required for this study has to be collected from 2 different sources. First, we need Open, High, Low, Close (OHLC) data for the required stocks and we need the tweets from twitter to get the sentiment score.

* The Historical price data is obtained from Yahoo Finance.
* The tweets are obtained by scraping the Twitter website.

**DATA DESCRIPTION**

**Twitter Data of Reliance Stock**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Tweet | Location | Friends\_count | Favourites\_Count | Followers\_count | Hashtags |
| 01-01-2016 | hny to all | New Delhi, Delhi | 340 | 28420 | 1220 | ['IPO', 'stocks', 'FII', 'DII', 'RIL', 'Sensex', 'Nifty'] |
| 01-01-2016 | happy birthday to the one and only dolla bill much love unc | BBU Alum ðŸ˜ˆ | 1473 | 2011 | 1113 | ['RIL'] |
| 01-01-2016 | what have prompt to interview delhi cm and give massive coverage to his | Polska | 1213 | 742 | 31865 | ['FalsePropaganda', 'RIL', 'GasDeals', 'NDA', 'Penalty'] |
| 01-01-2016 | kept and price | New Delhi, Delhi | 189 | 0 | 240 | ['RIL', 'PE', 'PP', 'PVC', 'unchanged'] |
| 01-01-2016 | love this picture for some reason | Warszawa | 0 | 5 | 146 | ['StephenCurry', 'AyeshaCurry', 'DellCurry', 'SethCurry', 'SydelCurry', 'SonyaCurry', 'Ril'] |

**Twitter Data for TCS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Tweet | Location | Friends\_count | Favourites\_Count | Followers\_count | Hashtags |
| 01-01-2016 | yay bake mac and lechon thank you team panorama | East End, Long Island | 1635 | 2780 | 314 | ['Horizon', 'TCS', 'January1st', 'Holidayshift'] |
| 01-01-2016 | happy and healthy from be proud of your accomplishment and learn from your | PaKisTaN | 639 | 705 | 344 | ['TCS', 'setgoals'] |
| 01-01-2016 | you can now try on your android phone.  -  passport express service announce for home delivery of passport | India | 0 | 3 | 2866 | ['FireFox', 'TCS'] |
| 01-01-2016 | visit today | India | 343 | 71 | 62 | ['wipro', 'tcs', 'cognizant', 'infosys', 'igate', 'mphasis', 'zensar', 'patni', 'ibm'] |
| 01-01-2016 | tx.i aplied renew of my son on dec apt dec dispt on jan amaze govt be wrkng thx | Pune, India | 189 | 1898 | 84 | ['Passport', 'TCS', 'PSKmalad'] |

This is the Twitter Dataset. We have more than 300,000 tweets. Variables included in the dataset are described below.

Date: - Date of the tweet creation

Tweets: - Tweets related to that particular stock on the particular date

Location: - The location from where user tweeted

Friends\_Count :- Count of friends a user has

Favorites\_Count :- Count of favourites a user has

Followers\_Count :- Count of followers a user has

Hashtags: - Hashtags used in the tweet

**Stocks data for Reliance**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 01-01-2016 | 500.159 | 504.6663 | 499.3665 | 502.9079 | 483.5406 | 2499742 |
| 04-01-2016 | 497.7815 | 502.1402 | 488.7174 | 492.9771 | 473.9922 | 13923887 |
| 05-01-2016 | 495.4536 | 500.2581 | 493.8191 | 497.8558 | 478.683 | 6897687 |
| 06-01-2016 | 499.0693 | 514.3247 | 495.5031 | 511.2538 | 491.565 | 12349673 |
| 07-01-2016 | 505.7312 | 509.1736 | 499.2922 | 501.8678 | 482.5405 | 9109980 |

**Stocks data for TCS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 01-01-2016 | 1219.5 | 1219.5 | 1206.125 | 1208.2 | 1062.938 | 712262 |
| 04-01-2016 | 1205.075 | 1207 | 1183.025 | 1184.8 | 1042.352 | 1870184 |
| 05-01-2016 | 1192.5 | 1193.3 | 1170.5 | 1174.475 | 1033.268 | 2678020 |
| 06-01-2016 | 1175.1 | 1193.075 | 1175.1 | 1190.8 | 1047.63 | 2653228 |
| 07-01-2016 | 1185 | 1191.45 | 1180 | 1185.625 | 1043.077 | 3199580 |

This is a stock market dataset. In our dataset we have 982 rows. Variables included in the dataset are described below.

Date: - Date of the given day

Open: - The price at which a security first trades when an exchange opens for the day

Close: - It is is referred to the price of a stock at the end of the trading hours

High: -The highest price for a stock on that particular day

Low: - The lowest price for a stock on that particular day

Adj. Close: - Adjusted closing price refers to the price of the stock after paying off the dividends

Volume: - The total number of shares bought or sold on that particular day

**RESEARCH METHODOLOGY**

Research methodology is the specific procedure or technique used to identify, select, process, and analyse information about the topic at hand. For this project, the steps are as follows -

1. Data Collection from Twitter and Yahoo Finance takes place and it is stored in a

database.

1. Data Cleaning and Pre-processing is done as per the points mentioned in detail in the

next section.

1. Fit a suitable Time Series model on the Historical prices only. Here, we fitted the

Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model due to the volatility in data.

1. After this, we fit Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM)

on 4 years of historical price data to get the baseline model.

1. We run Sentiment Analysis on the tweets data and get the sentiment scores of all

tweets and group them by day to get a single value of sentiment score.

1. Combine the historical price and sentiment score into a single file and fit the models.

GRU and LSTM are fitted and the predictions are calculated.

1. Analysis and Interpretation of the results obtained in testing of the fitted models.

**DATA PREPROCESSING**

Twitter Sentiment Analysis

We built a web scraper using SNScrape. The data cleaning process includes -

* The first step in pre-processing includes removing unicodes from the tweets column.

For ex: - 0 is denoted as U+0000 , 1 is denoted as U+0001. If data is present in such a format Unicodes are removed.

* The second step includes replacing URLs and the username beginning with ‘@’

present in the tweets as they are of no use.

* All the punctuations including full stop and question mark are removed and the whole

string is converted into lowercase.

* The new line characters “\n” are removed along with the numbers, HTML tags

removal and multiple space removal.

* Emojis were converted to words using the emot library and contractions were

replaced by full words. e.g.: ain’t = am not, can’t = cannot, etc.

* Since the tweets were in different languages, we used GoogleTranslator to convert

them into English.

* Finally, we go for stopword removal, lemmatization and Part-of-Speech tagging using

wordnet.

Historical Price Data

Fourier analysis transforms a signal from the domain of the given data, usually being time or space, and transforms it into a representation of frequency. In the financial field the FFT (Fast Fourier Transformation) is used in computational finance mostly, for predicting the prices of financial derivatives on a time series basis.

* We get the technical indicators like Moving Average Convergence Divergence

(MACD), 7-day and 21-day Moving Average, Bollinger Bands, Exponential Moving Average and Log-momentum.

* Fourier Transformation with 3, 6 and 9 components are done and the output is used as

an input in modelling.

* Check for missing values and use bfill() for imputation.
* Plot the Autocorrelation plot to check for the correlations present in past lags.
* Normalise the data using MinMaxScaler and reshape it so that 30 days input is used

for 1 day prediction.

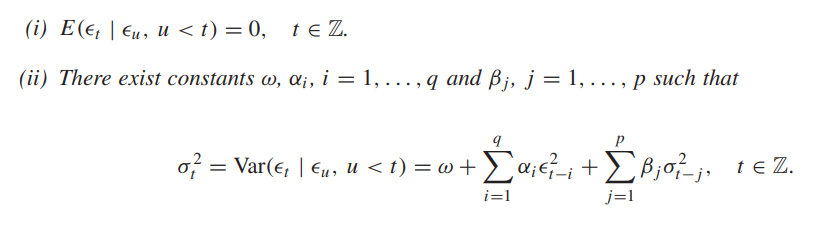
**TECHNIQUES USED**

**ARCH AND GARCH**

The ARCH and GARCH models, which stand for autoregressive conditional heteroskedasticity and generalised autoregressive conditional heteroskedasticity, are designed to deal with just this set of issues. They have become widespread tools for dealing with time series heteroskedastic models.

An ARCH (autoregressive conditionally heteroscedastic) model is a model for the variance of a time series. ARCH models are used to describe a changing, possibly volatile variance. Although an ARCH model could possibly be used to describe a gradually increasing variance over time, most often it is used in situations in which there may be short periods of increased variation.

A GARCH (generalised autoregressive conditionally heteroscedastic) model uses values of the past squared observations and past variances to model the variance at time t. A process (t) is called a GARCH(p, q) process if its first two conditional moments exist and satisfy:



Checking whether ARCH effect is present or not in our data.

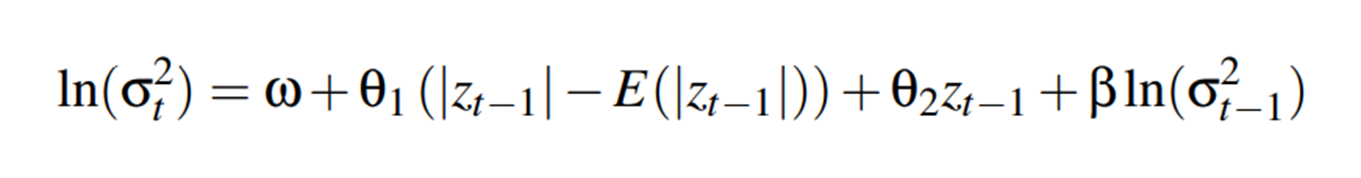
H0 : The ARCH effect is not present.

H1 : The ARCH effect is present.

**eGARCH model: -**

To overcome the drawbacks of GARCH model eGARCH model is used which is Exponential GARCH model.

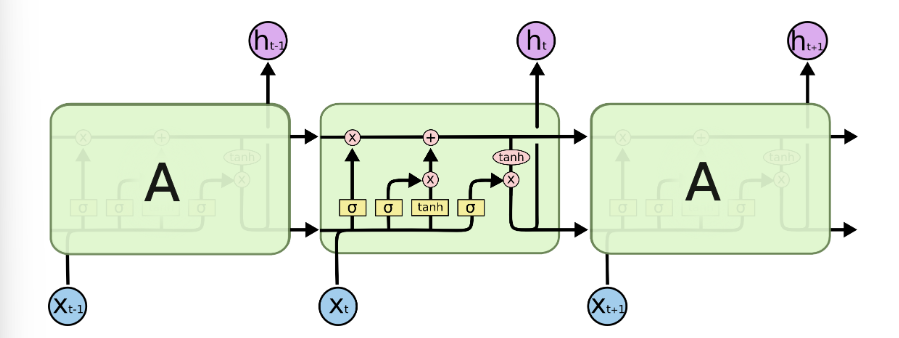
The eGARCH (1,1) model is given by



This model takes into account the asymmetric effect of positive shock and negative shocks.

**LSTM**

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour. LSTMs have this chain-like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.



The repeating module in an LSTM contains four interacting layers.

The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.

The cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It’s very easy for information to just flow along it unchanged.

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

An LSTM has three of these gates, to protect and control the cell state. These cells use the gates to regulate the information to be kept or discarded at loop operation before passing on the long term and short term information to the next cell. We can imagine these gates as Filters that remove unwanted selected and irrelevant information. There are a total of three gates that LSTM uses as Input Gate, Forget Gate, and Output Gate.

Input Gate

The input gate decides what information will be stored in long term memory. It only works with the information from the current input and short term memory from the previous step. At this gate, it filters out the information from variables that are not useful.

Forget Gate

The forget decides which information from long term memory be kept or discarded and this is done by multiplying the incoming long term memory by a forget vector generated by the current input and incoming short memory.

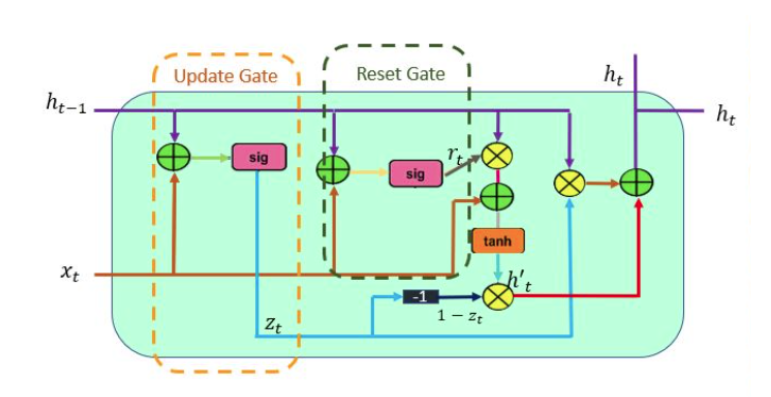
Output Gate

The output gate will take the current input, the previous short term memory and newly computed long term memory to produce new short term memory which will be passed on to the cell in the next time step. The output of the current time step can also be drawn from this hidden state.

So, this is all about the mechanism of LSTM to realise this with practical implementation. Here I have demonstrated the LSTM use case in which you can check input and output sequences with their shape.

**GRU**

The workflow of the Gated Recurrent Unit, in short GRU, is the same as the RNN but the difference is in the operation and gates associated with each GRU unit. To solve the problem faced by standard RNN, GRU incorporates the two gate operating mechanisms called Update gate and Reset gate.



Update gate

The update gate is responsible for determining the amount of previous information that needs to pass along the next state. This is really powerful because the model can decide to copy all the information from the past and eliminate the risk of vanishing gradient.

Reset gate

The reset gate is used from the model to decide how much of the past information is needed to neglect; in short, it decides whether the previous cell state is important or not.

First, the reset gate comes into action it stores relevant information from the past time step into new memory content. Then it multiplies the input vector and hidden state with their weights. Next, it calculates element-wise multiplication between the reset gate and previously hidden state multiple. After summing up the above steps the non-linear activation function is applied and the next sequence is generated. This is all about the operation of GRU, the practical examples are included in the notebooks.

**MODEL FINDINGS**

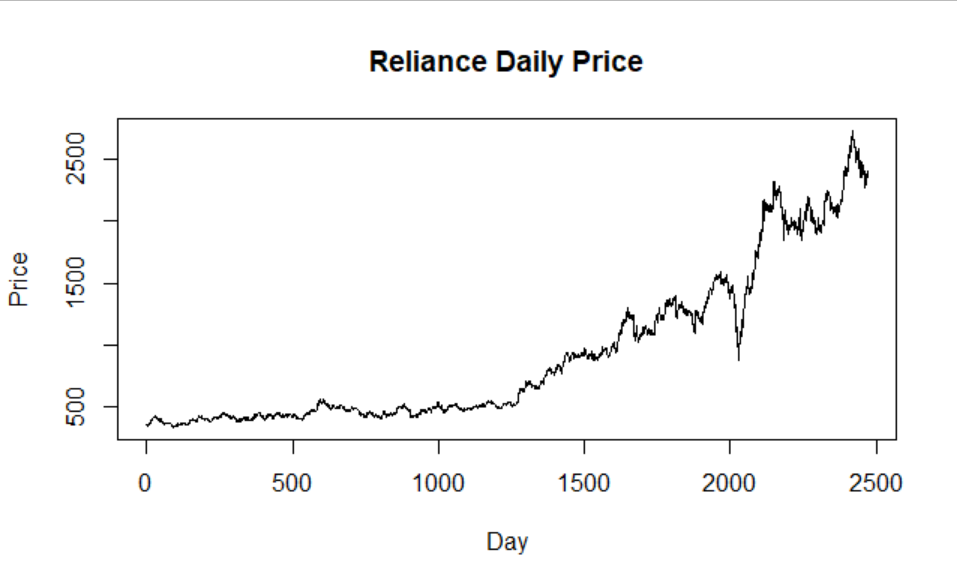
After completing the Data Pre-processing, the next step is fitting the suitable models and predicting the future price of the stocks in concern and validating their accuracy.

For the purpose of this project, we have fit GARCH, GRU and LSTM models with specific parameters. The forecasts and the valuations are shown below in a chronological manner.

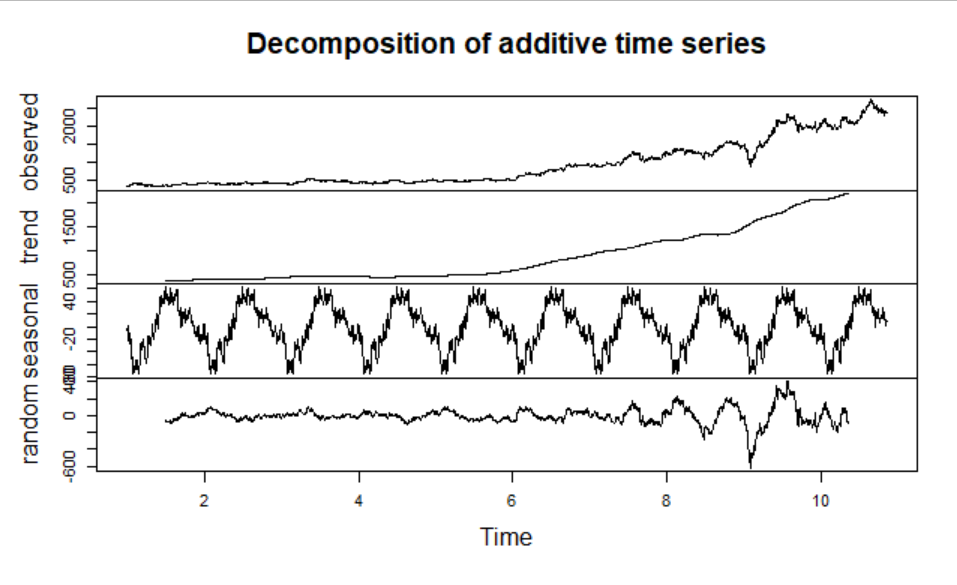
**FOR RELIANCE**

1. **GARCH (1,1)**

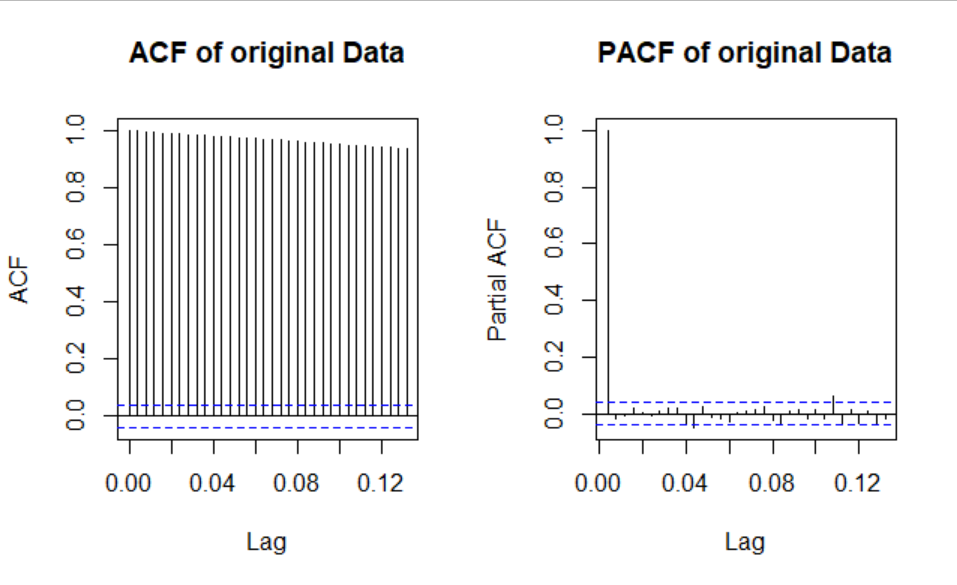
The following plot shows daily prices for reliance stocks.



The decomposition plot for reliance stock is given below.

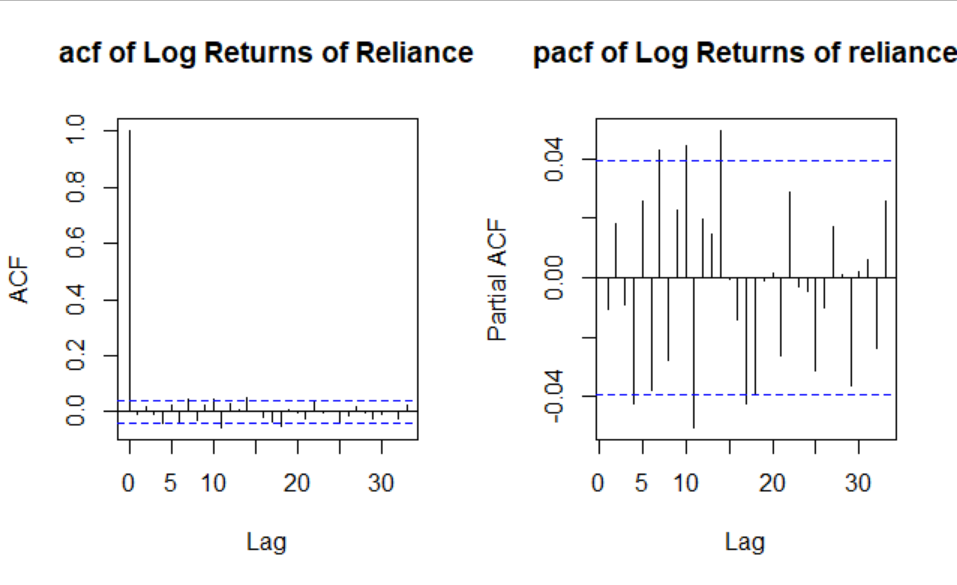


From this plot, we can conclude that there is a trend and seasonality is present in the data.

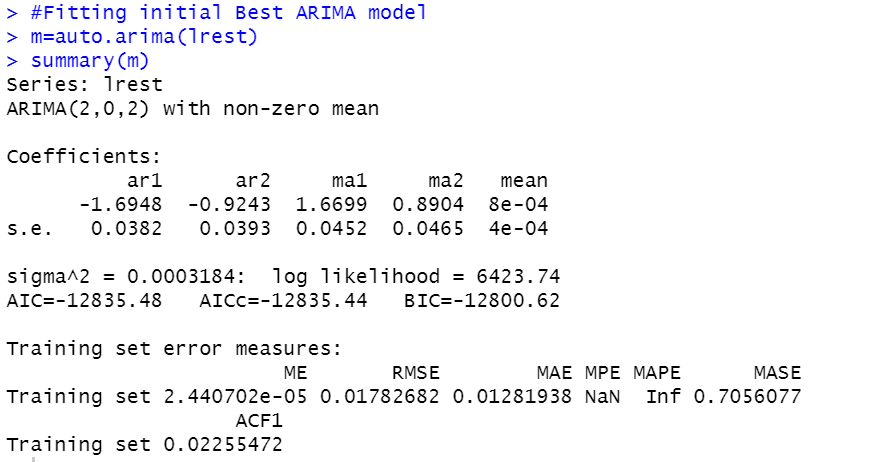


From the ACF and PACF plots we clearly see that the data is not stationary.

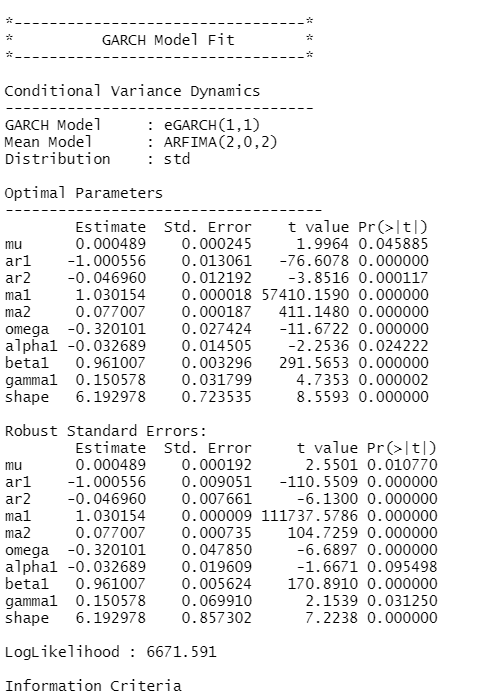
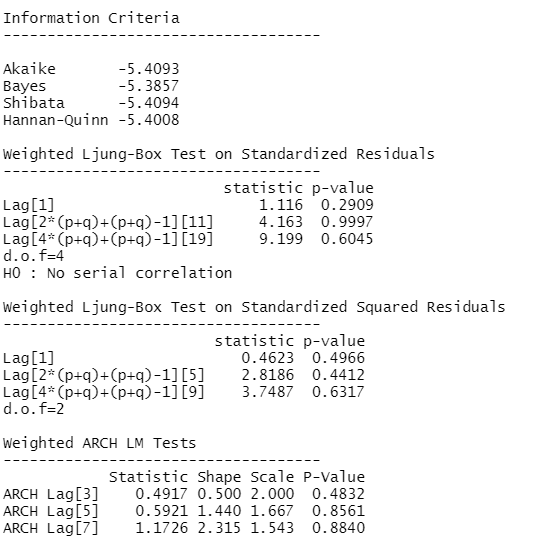
Next to make the data stationary we will calculate the log returns and plot the ACF and PACF of the log returns.

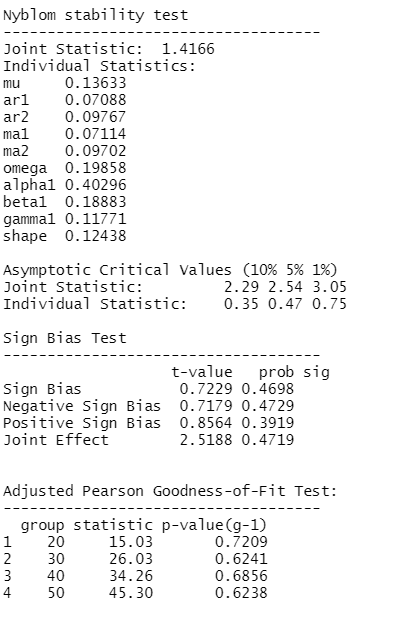


**Fitting Initial Best ARIMA Model**



The initial ARIMA model we obtained using auto Arima is ARIMA (2,0,2).

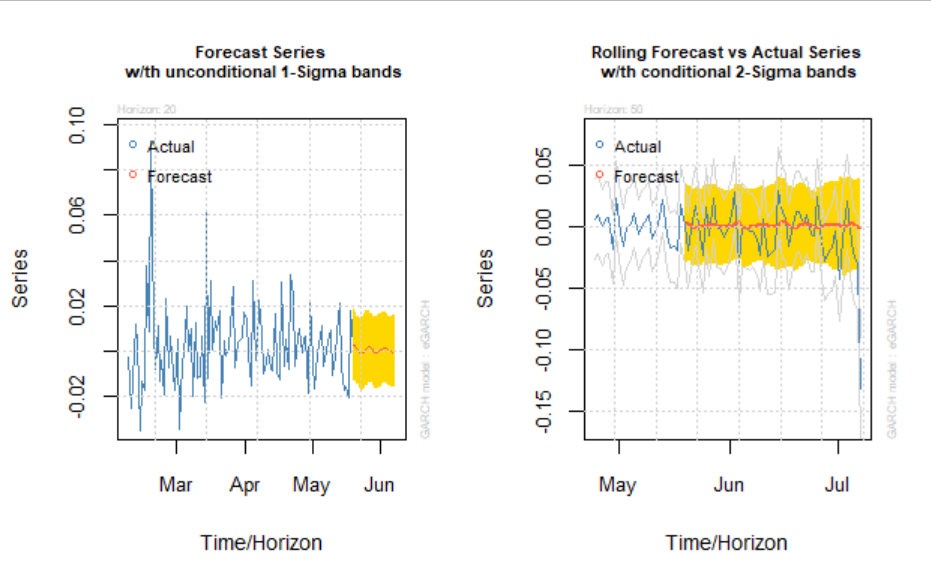
 

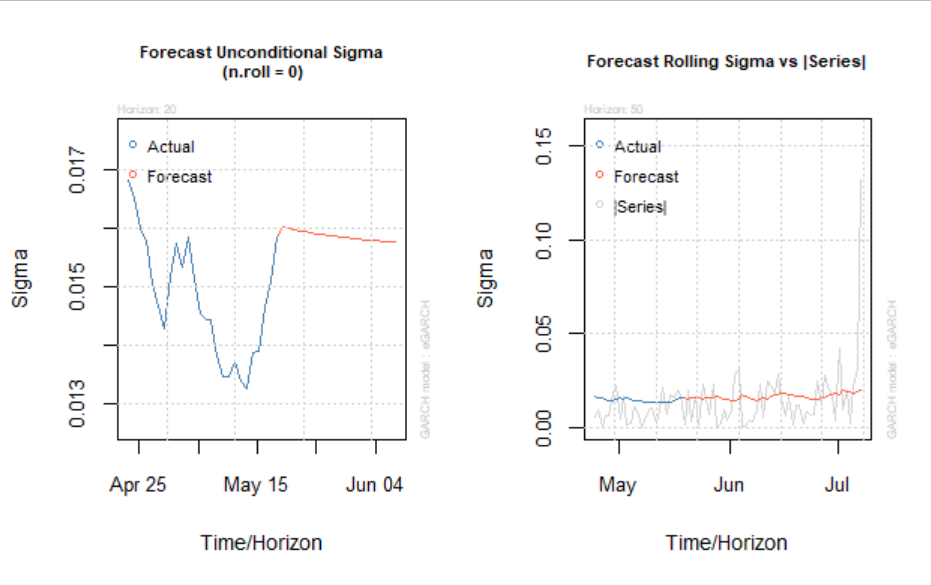


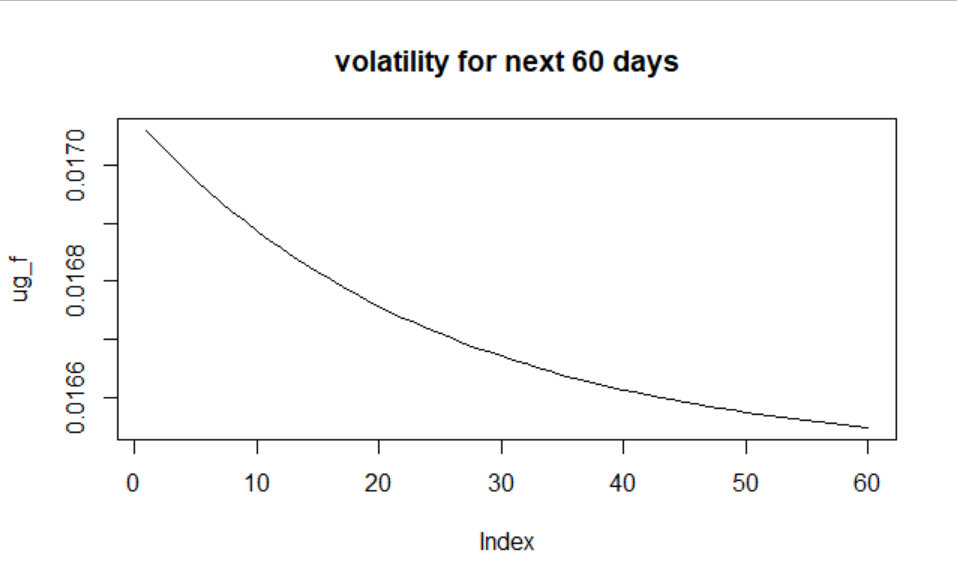


The GARCH (1,1) model is producing the above results.

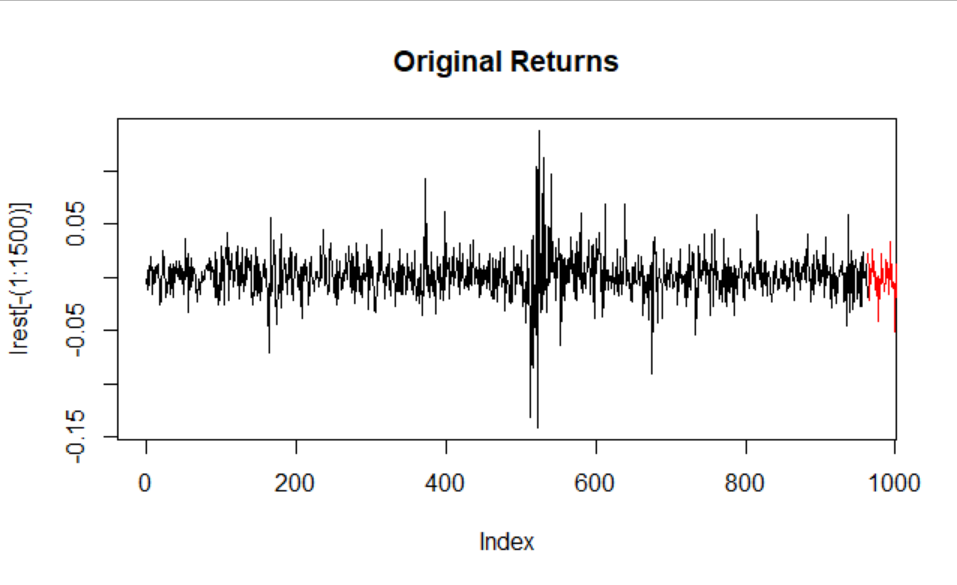
Then, the forecast was made using the GARCH (1,1) model which can be visualised below.







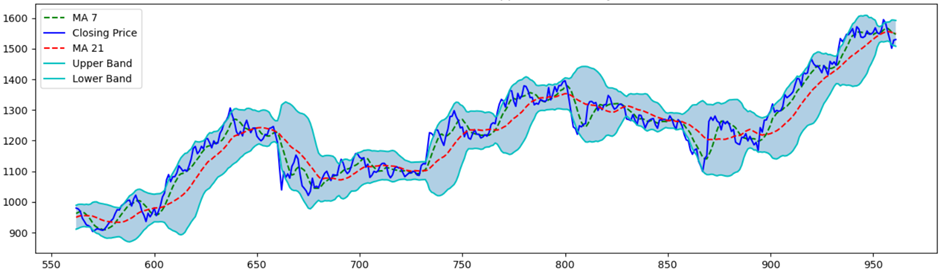
Comparing the prediction with original data we can see that volatility was relatively less than previous times.



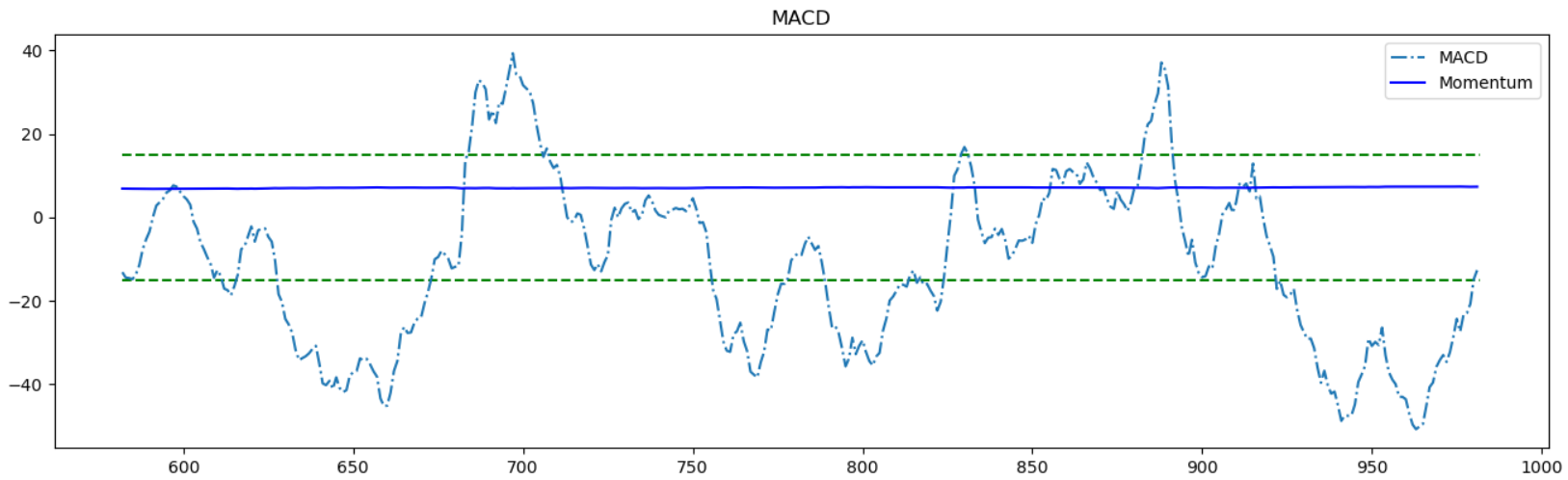
1. **Gated Recurrent Units (GRU)**

2 GRU models are fit. One only with the historical price data and one with the historical price data combined with sentiment scores.

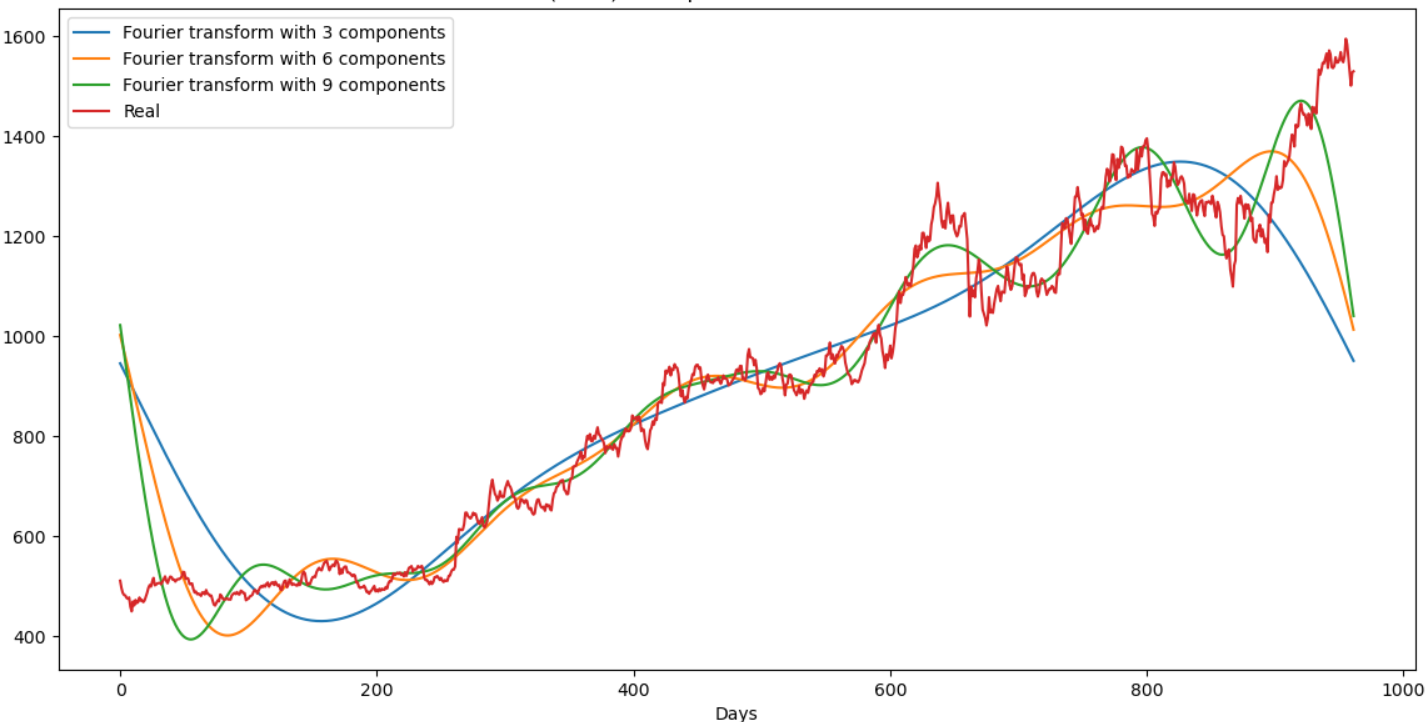
The 400-day indicator graph is shown below.



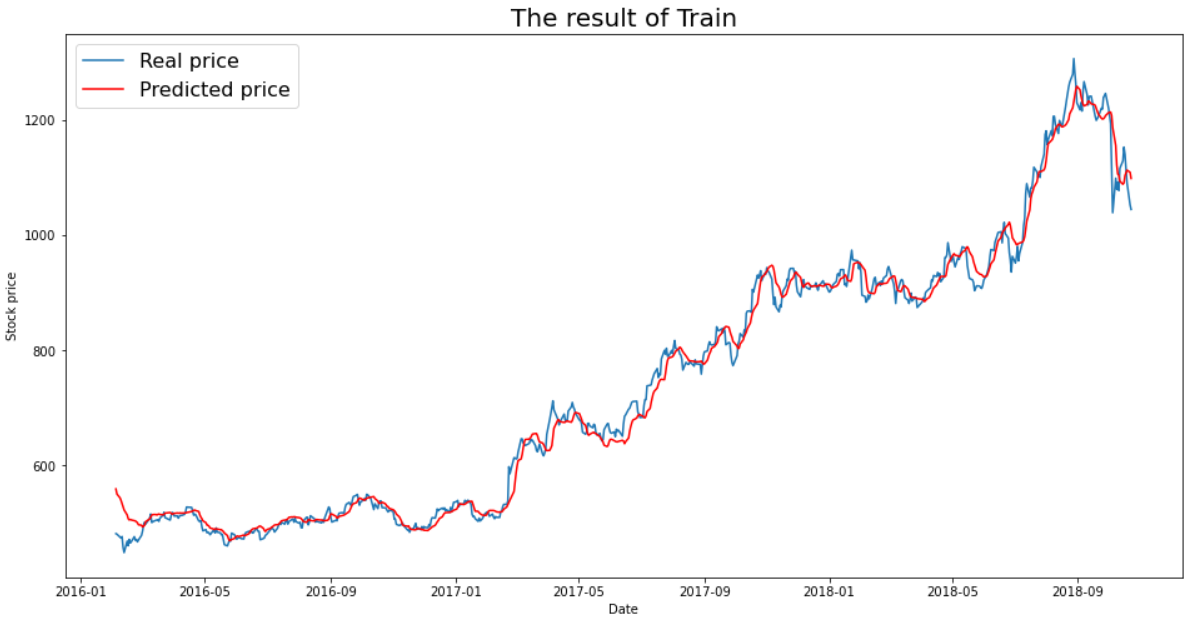
The Moving Average Convergence /Divergence (MACD) is shown below.



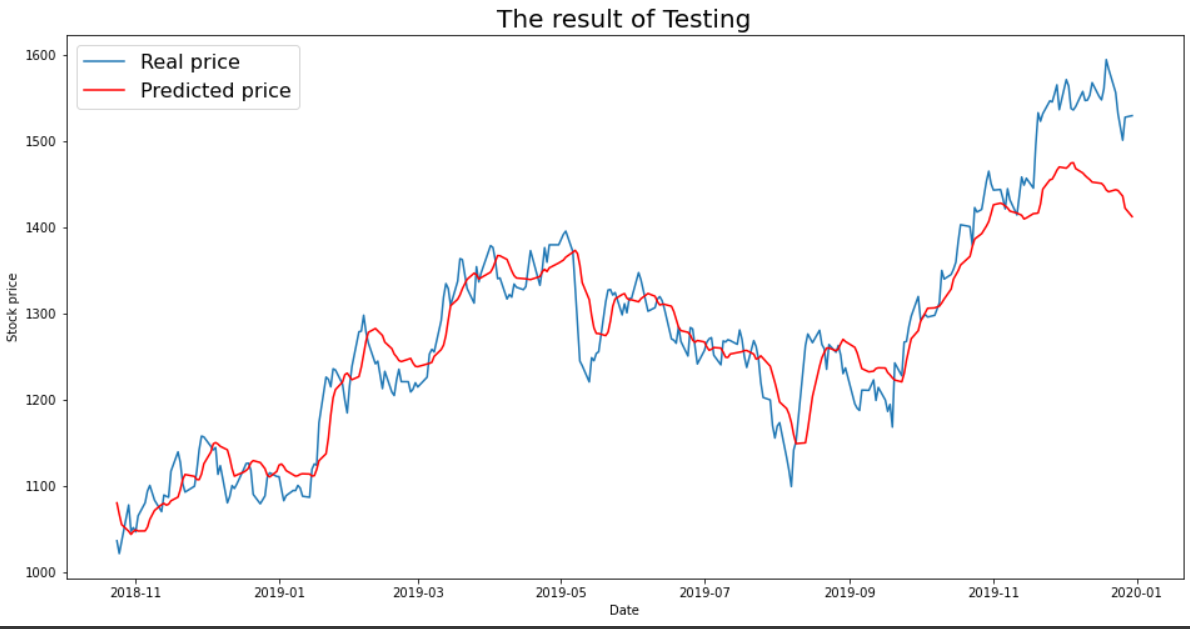
Fourier Transformation is done and the graphical representation with 3, 6 and 9 components are done.



The training and testing forecasts for GRU (Basic) are shown below.



RMSE= 22.83

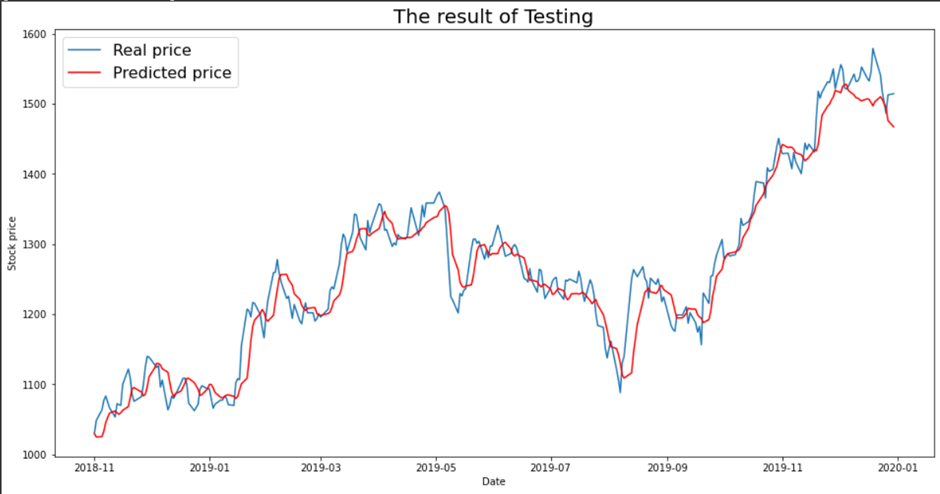


RMSE = 44.55

The training and testing forecasts for GRU (with sentiment scores) are shown below.



RMSE = 22.09



RMSE = 58.47

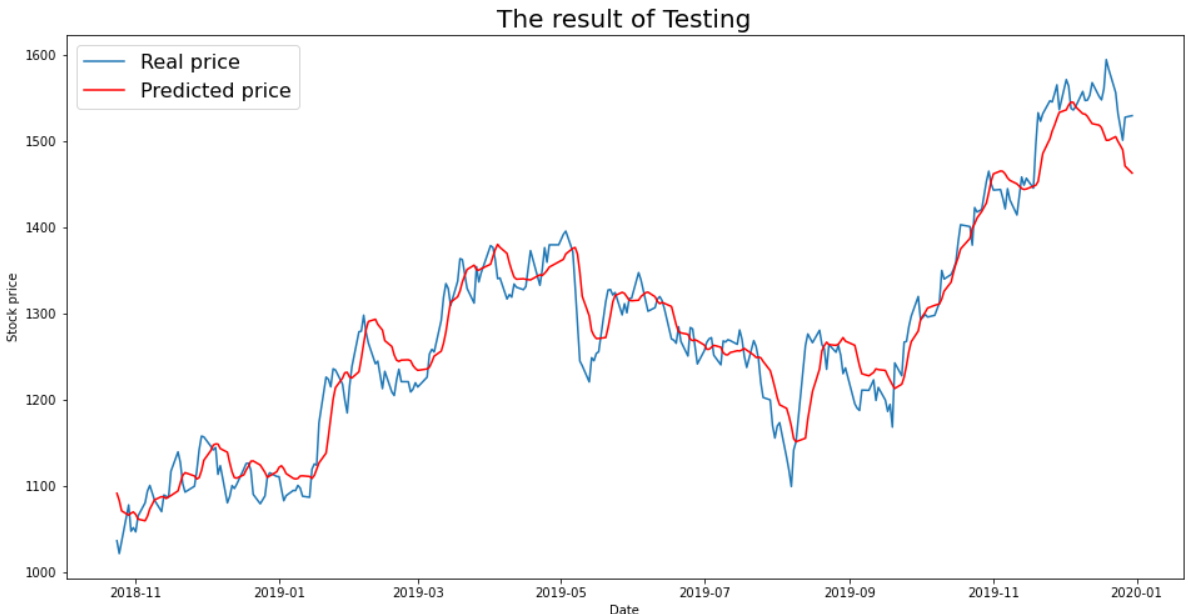
1. **Long Short Term Memory (LSTM)**

As for the GRU, we fit 2 models for LSTM as well, one with sentiment scores and one without the scores.

Basic LSTM



RMSE = 52.63

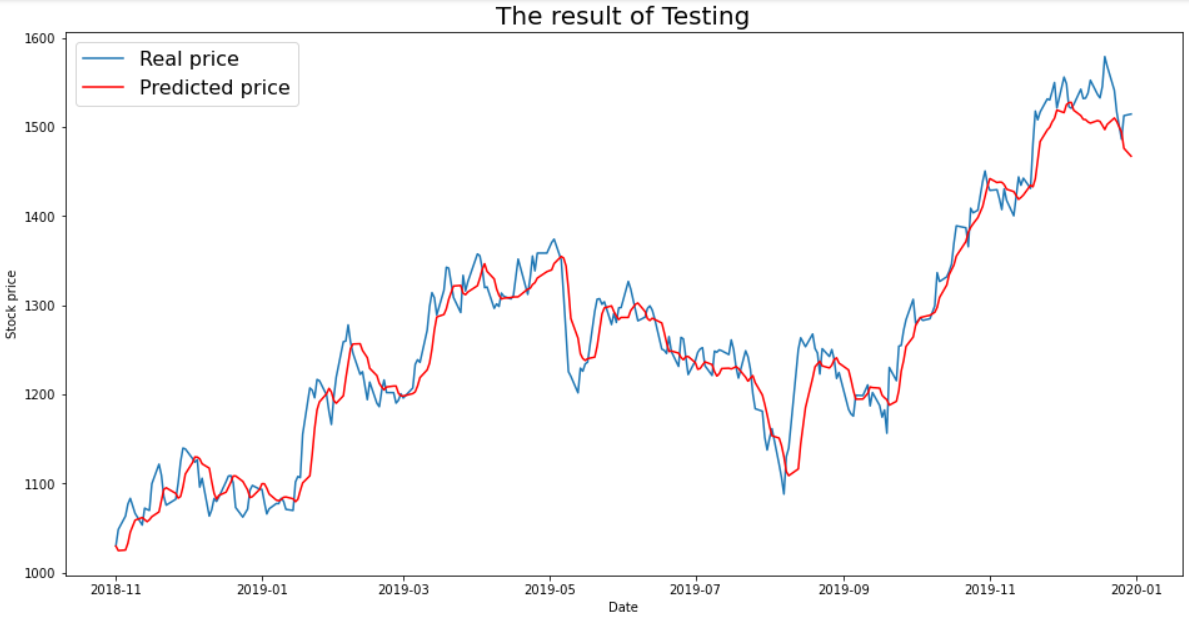


RMSE = 32.80

The training and testing forecasts for LSTM (with sentiment scores) are shown below.



RMSE = 19.20

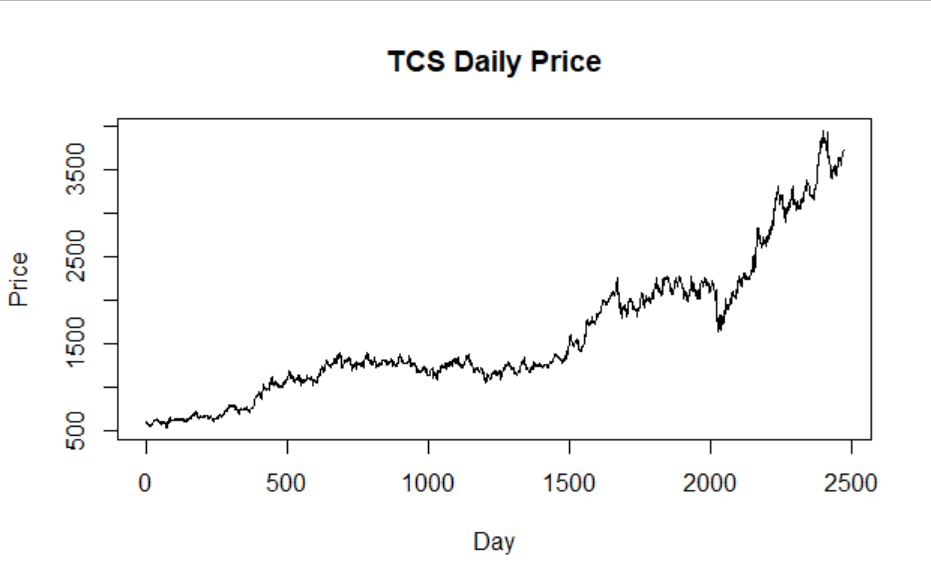


RMSE = 30.58

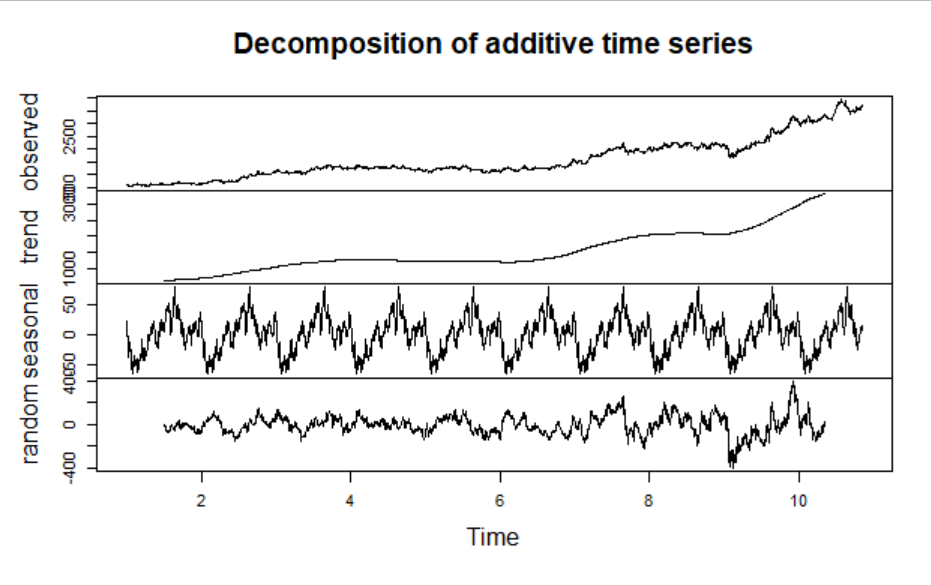
**TATA CONSULTANCY SERVICES(TCS)**

1. **GARCH (1,1)**

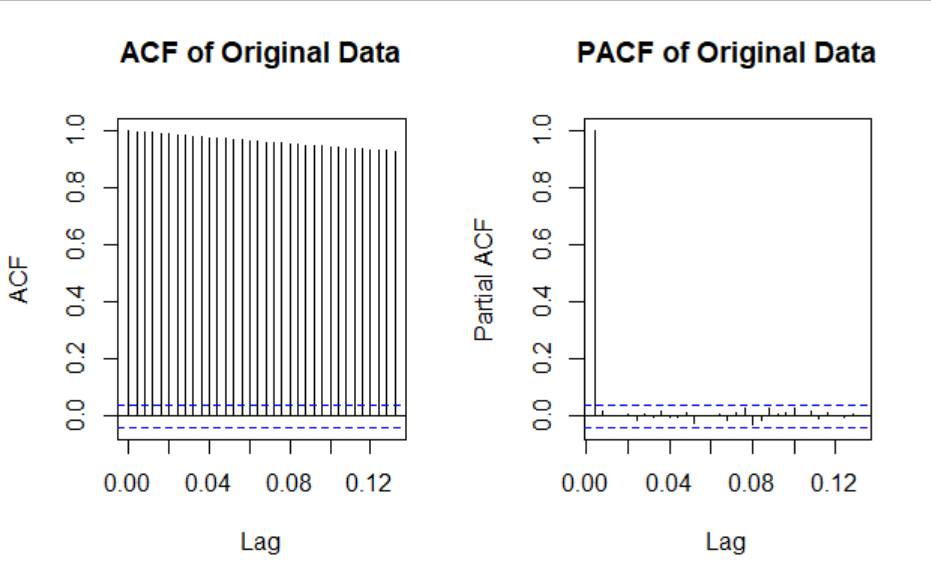
Graph for the daily closing price of TCS



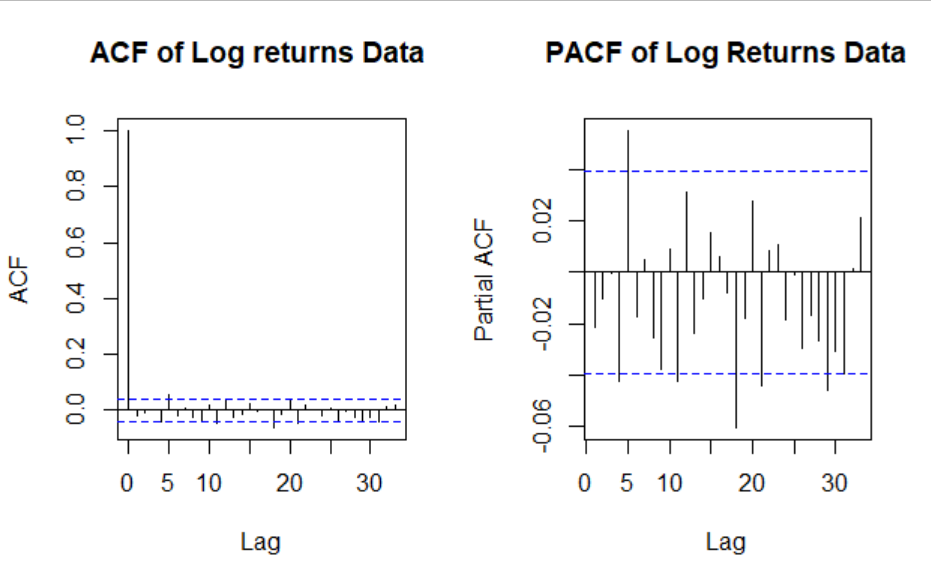
Plotting ACF and PACF of above data



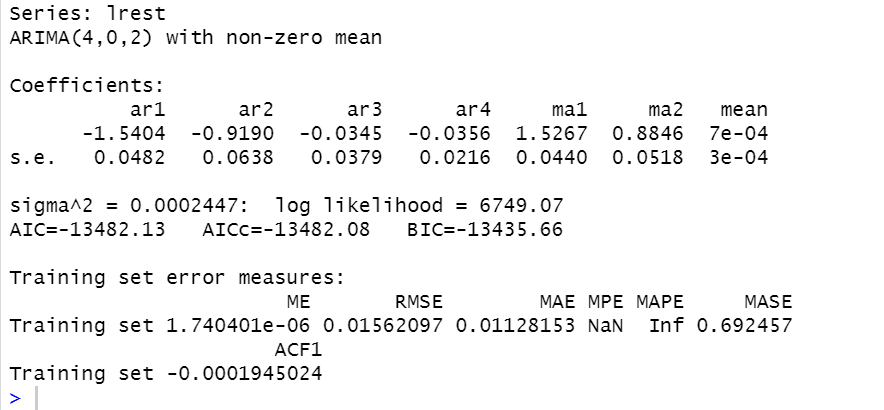
From the above plot we can clearly see that trend and seasonality is present in the data.



The data is not stationary.

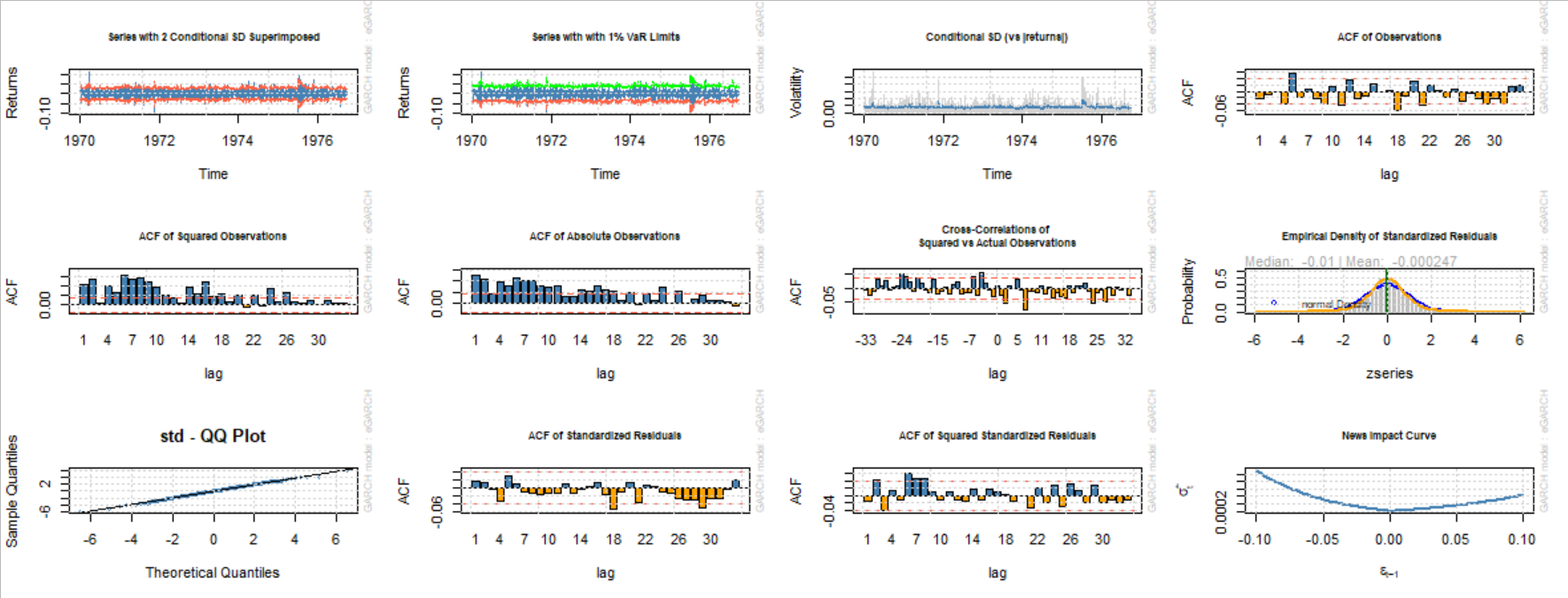


Checking for ARCH effect in our log returns data,

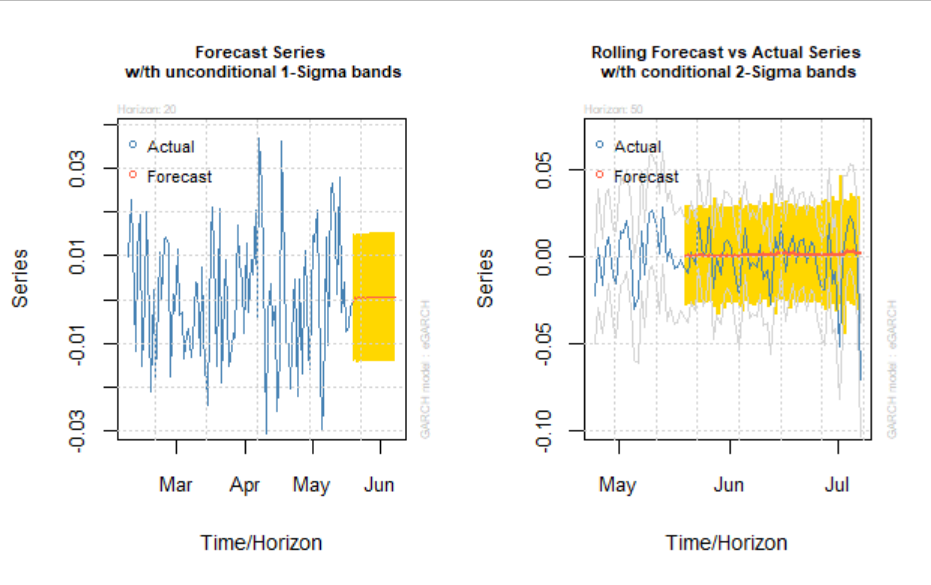


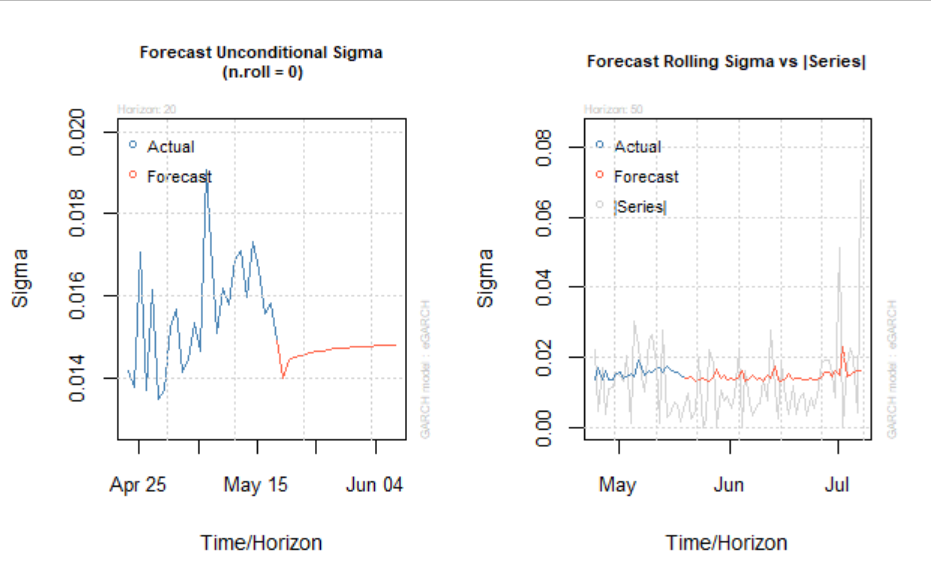
The initial best ARIMA model is ARIMA (4,0,2)

After fitting GARCH (2,1) we get



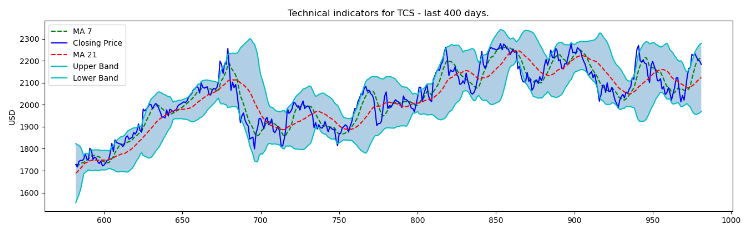
The forecast was made and can be shown as



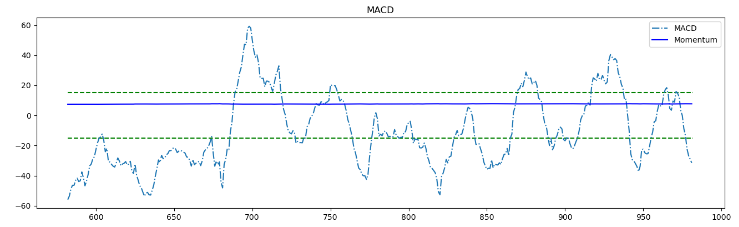


1. **Gated Recurrent Units (GRU)**

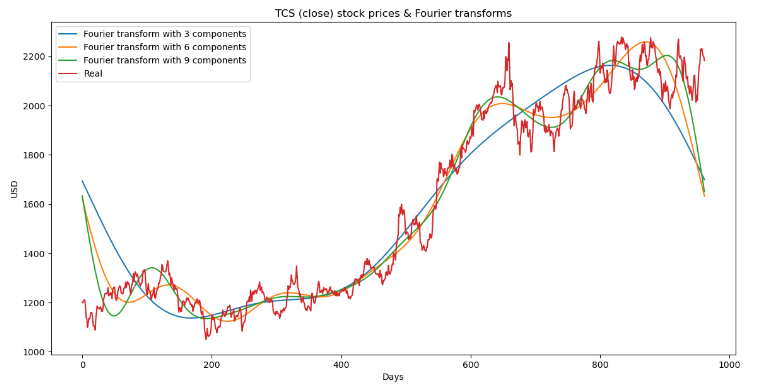
The 400-day indicator graph is shown below.



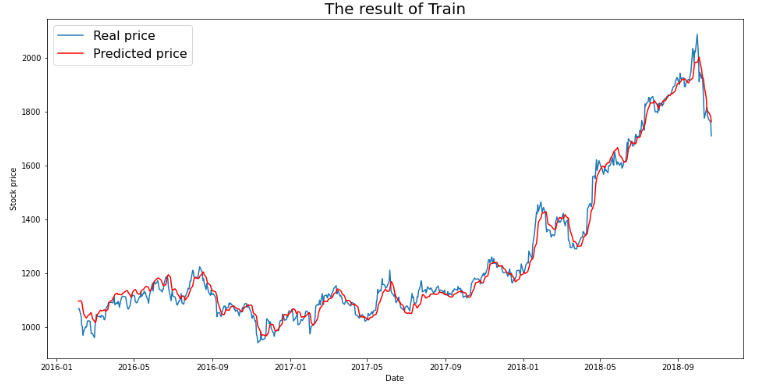
The Moving Average Convergence /Divergence (MACD) is shown below.



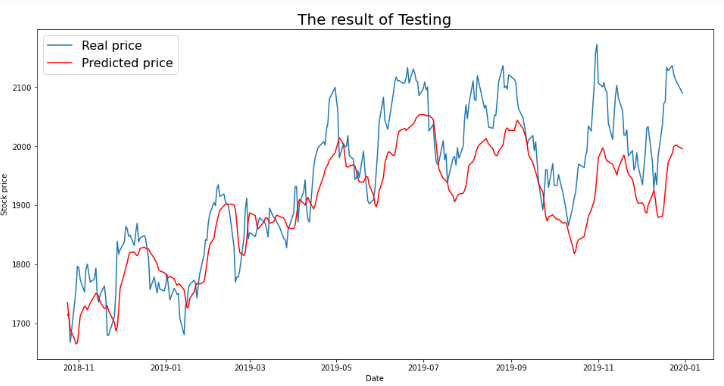
Fourier Transformation is done and the graphical representation with 3, 6 and 9 components are done.



The training and testing forecasts for GRU (Basic) are shown below.



RMSE = 32.3381

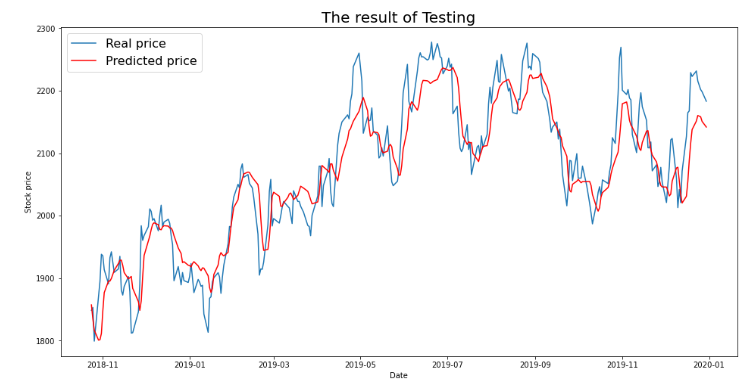


RMSE = 61.96

The training and testing forecasts for LSTM (Basic) are shown below.

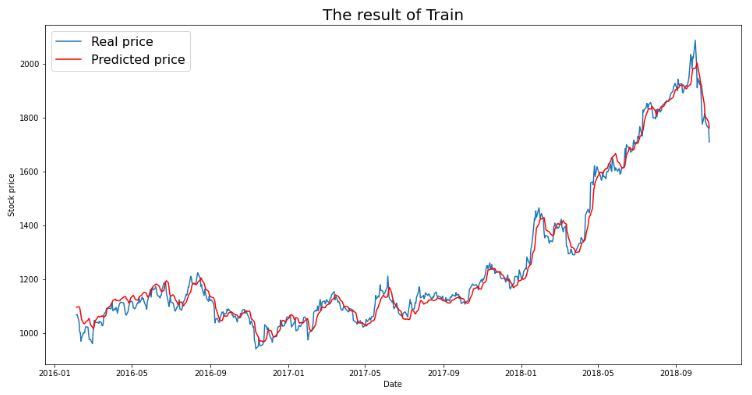


RMSE = 25.01

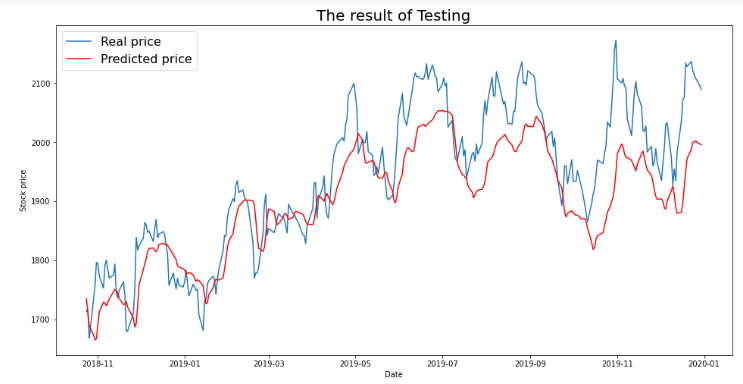


RMSE = 44.42

The training and testing forecasts for GRU (with sentiment scores) are shown below.



RMSE = 31.0119

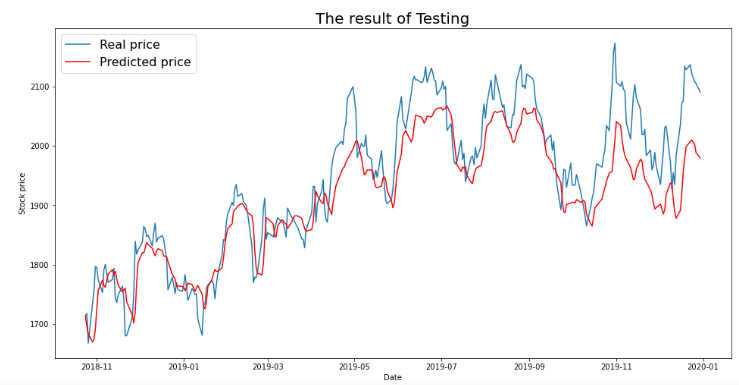


RMSE = 72.2782

The training and testing forecasts for LSTM (with sentiment scores) are shown below.



RMSE = 23.9379



RMSE = 56.9948

**CONCLUSION**

* The Garch Model predicts the volatility of the stock prices and forecasts the future price with high accuracy.
* There is a positive relationship between stock price and twitter sentiment.
* Price prediction using the combined approach of Sentiment Analysis yields a lower RMSE by almost 10%.

**FUTURE SCOPE**

For the purpose of expanding the horizon of the study, one can

* Get the news headlines daily for additional information.
* Use word embedding models while performing sentiment analysis.
* Take multiple stocks and do multivariate analysis.
* Fit a deep learning model with more hidden layers.
* Increase the training data size.

**REFERENCES**

[1] Hsiao-Chuan Chou, Kandethody M. Ramachandran. “Combining Time Series and Sentiment Analysis for Stock Market.” *3rd International Conference on Statistics: Theory and Applications (ICSTA'21)*. Virtual Conference: Virtual Conference – July, 2021, 2021. 11.

[2] Anika Kanwal, Man Fai Lau, Sebastian P.H. Ng, Kwan Yong Sim, Siva Chandrasekaran. “BiCuDNNLSTM-1dCNN — A hybrid deep learning-based predictive model for stock price prediction.” *Science Direct* (2022): 15.

[3] Olivier Kraaijeveld, Johannes De Smedt. “The predictive power of public Twitter sentiment for forecasting cryptocurrency prices.” *Science Direct* (2020): 22.

[4] Nan Jing, Zhao Wu, Hefei Wang. “A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction.” *Science Direct* (2021): 12.

[5] Kevin Hu, Daniella Grimberg, Eziz Durdyev. “Twitter Sentiment Analysis for Predicting Stock Price Movements.” 2021.

[6] Anshul Mittal, Arpit Goel. “Stock Prediction Using Twitter Sentiment Analysis.” 2021.

**GLOSSARY AND TERMINOLOGIES**

Sentiment Analysis: - The process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Stop Words: - Stop words are the words in a stop list (or *stoplist* or *negative dictionary*) which are filtered out (i.e., stopped) before or after [processing of natural language](https://en.wikipedia.org/wiki/Natural_language_processing) data (text) because they are insignificant.

Stemming: - Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the pre-processing of text, words, and documents for text normalisation.

Lemmatization: - It usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

Bag Of Words: - When dealing with text corpus we come across multiple words which we use in Natural Language Processing (NLP) applications to get meaningful insight. To do that we convert those words into something which the model can understand using a concept called “Bag of Words”.

Time Series: - A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.