Abstract--

Today, the stock market is attracting a lot of attention from various big Investment firms through its high-level risks and the high returns involved with it. Stock Exchange markets typically involve securities investments like NYSE, NSE, DJIA, etc. that are advantageous to the global economy. A novel approach for prediction of Stock Market fluctuations is introduced through opinion mining by classification of sentiments of Twitter messages related to that stock. Technology industries like Apple Inc were focused. Other industries related to healthcare and other fields can also be used in the proposed approach. The future stock returns like Closing Price and Daily Change, have some predictive properties with respect to publicly available information of present sentiments on Social Media like Twitter and historical stock market datasets on NASDAQ, Yahoo! Finance, or Google Finance. A prediction model based on the analytics for forecasting the movements of Stock markets for next day is proposed in this paper through sentiment analysis using NLP and time series forecasting for a short-term prediction like ARIMA model and Neural Network for finding hidden states. One approach was related to Analysis of current sentiments related to a firm like that of Apple Inc. from social media and news. The other approach involved using Time Series Forecasting methods on previous historical data to find patterns in short-term seasonal intervals predict the most likely bounded price change and its accuracy. The most probabilistic frequency of stationary deviations in closing price, found from the Ensemble model of ARIMA and Neural Network, were predicted through Random Walk usage on 40 years stock prices of Apple Inc. on a Hadoop cluster. It will help in automating the forecasting of stock price movements in future and gives a helping hand for global financiers, angel investors and investment bankers to decide the right time for selling and purchasing of stocks in correlation with public sentiments. The results are shown in terms of sentiment analysis outputs and visualizations using R, Python, Java, and Hadoop. The correlations and forecasts obtained conclude the proposed model’s good potential for prediction of stock market movements on a short-term basis and consistent with public sentiments.

Keywords— Twitter Sentiment Analysis Opinion Mining, Time Series Forecasting, Stock Market Investment Strategies, Prediction Algorithms, Big Data, Ensemble Models

Abbreviations—

ACF Auto Correlation Function

PACF Probabilistic Auto Correlation Function

AIC: Akaike Information criteria

AICc Corrected Akaike Information criteria

BIC Bayesian Information Criteria

AR Autoregressive Model

ARIMA Auto Regressive Integrated Moving Average Model

DJIA Dow Jones Industrial Average

ETS Exponential Timeseries Smoothening

MA Moving Average

MAPE Mean Average Percentage Error

ML Machine Learning

NASDAQ National Association of Securities Dealers Automated Quotations

NLP Natural Language Processing

NSE National Stock Exchange

NYSE New York Stock Exchange

RMSE Root Mean Square Error

Notations

X[t] the prediction of Deviation in current day’s closing price and returns

u the most probabilistic frequent boundary value of drift in Random Walk

p magnitude of Probabilistic logarithmic Regression in AR model

d magnitude of differencing in Integrated model

q number of combination of lagged error terms in MA model

L logarithmic residuals in ARIMA forecasts

i value of training interval in time series

y(v[i]) output of the ith node (neuron) in neural network

v[i] weighted sum of the input connections inside hidden layer

w[I,j] each node in one layer of neural network connects with a certain weight w[i,j] to every node in the following layers

Chapter-1 Introduction

1.1 General

In financial stock markets, it is considered to be an impossible task to predict most-likely stock market movement. There are generally two approaches for predicting the future market movements. One approach [1] relates to Analysis of current sentiments related to a firm like that of Apple Inc. from social media and news. The other approach [7] involves using Time Series Forecasting methods on previous historical data to find patterns in short-term seasonal intervals predict the most likely bounded price change and its accuracy.

1.2 Sentiment Analysis Approach

In the first approach of the proposed model, big data from social media sites like Twitter is used for analysis. Positive and negative sentiments are identified for companies using Natural Language Processing through word dictionaries and n-grams model. Public sentiments and opinions regarding Financial data news and currents events in Twitter are also streaming at an exceptional speed that must be dealt with in a well-timed manner to get the present thoughts of people regarding the Firm and its stocks and securities, which has also attracted the attention of a lot of researchers. When someone buys a stock, he/she generally buys it because of something he/she heard about on the news, social media or through their friends. This external information is used for stock market prediction. Based on this a correlation is found between Daily Price Change and sentiment score to predict next day's most likely movement.

1.3 Time Series Forecasting Approach

For the second approach, Historical stock market changes dataset is used for forecasting by assuming that certain patterns have bearings on the future for short-term linear intervals. This is what the ARIMA prediction model for time series forecasting is based on and is famous for. The stationary time series was used to forecast the closing price and returns in logarithms based on training of historical data and combined with Neural Network to get accuracy of 87%. The difference in ARIMA suggested the presence of Random Walk in the stock prices for short-term forecasting. This is the basis for another popular model known as Random Walk. The most probabilistic frequency of stationary deviations in closing price, found from the Ensemble model of ARIMA and Neural Network, were predicted through Random Walk usage on 40 years stock prices of Apple Inc. on a Hadoop cluster. These approaches can be combined with the prediction of Stock returns to uncover market patterns, plan strategies for future investments, and forecast the value of next day's closing price.

Chapter-2 Review of Literature

2.1 Research in forecasting of market movements through Twitter Sentiment Analysis

The most popular research published in prediction through sentiment analysis area is by Bollen’s [2] thorough investigation of the public sentiments into 6 dimensions of anxiousness, calmness, vitalness, kindness, sureness, and happiness obtained from feeds of Twitter data and its correlation with the Dow Jones Industrial Average (DJIA) Index values, found through usage of a Fuzzy Neural Network and Granger Causality Analysis for predicting and proving the existing correlation with an accuracy of 86.7 percent in market movements and closing values with very low MAPE (Mean Average Percentage Error). Chen and Lazer [8] built a prediction model to derive the investment strategies through the observation and classification of the twitter feeds into Financial positive and negative sentiments. On the basis of industry type in which people work, Bing et al. [9] studied twitter dataset to predict prices of stock markets. Some of the industries they used were Technology, Finance, Information Technology, etc. A strong negative correlation pattern was found out by Zhang [10], between public opinions of worriedness, fearfulness, and hopefulness in tweets and values of DJIA indices. A research in this field was also done by Brian et al. [11] recently, based on the correlation of stock price rise and fall with the public sentiments through the study of Pearson correlation coefficient for stocks. In a paper of Mittal et al. [12], it was proven that the mechanism of predicting with accuracy rate to be around 75 percent with a usage Fuzzy neural networks on DJIA Index and Twitter Feeds.

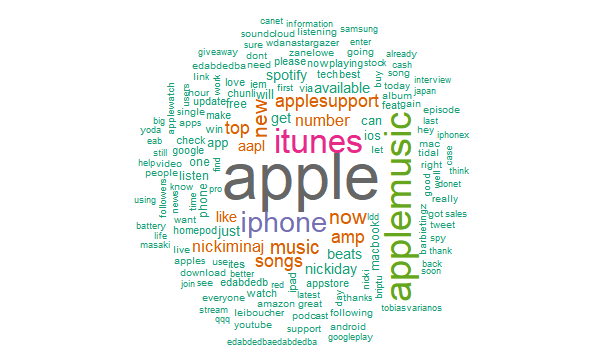
2.2 Research in predicting closing price from timeseries forecasting

Research in the prediction of Stock Markets has also been done with help of algorithms in Artificial Intelligence field by Bouktif et al. [17]. They combined Bayesian Classifiers with the Ant Colony Optimization algorithm on the public mood states obtained from Twitter giving a significant performance in prediction accuracy of concerned stock values. The approach related to Time series forecasting has also been researched through Deep Neural Networks by Gunduz et al. [18]. Convolutional Neural Network was used for prediction of the direction of movement of the closing price of GARAN, THYAO and ISCTR stocks. LSTM Neural Networks were also used in time series forecasts by Nelson et al. [20] for predicting future trends in the market through stock price history and analysis indicators with 56 percent accuracy of movement direction were found. An Ensemble model of forecasting has been studied by Bautu et al. [19] by ensembles of Gene Expression Programming (GEP) evolved models in deep learning and artificial intelligence fields through binary classification of stock prices. ARIMA model (Auto-Regressive Integrated Moving Average) was studied for forecasting in R by Angadi et al. [7] in depth and was found to be very accurate for short-term forecasting of stock market trends. EMMS (Expert Model Mining System) combining AR (Auto-Regression), I (Integration), MA (Moving Average) with ES (Exponential Smoothening) was researched by Rao et al. [1] reducing MAPE by combining it with social media Sentiment Analysis Correlations for NASDAQ (National Association of Securities Dealers Automated Quotations) and DJIA values with 75.56 percent accuracy through learning algorithms of Support Vector Machines and Neural Networks for Portfolio Management implemented in MATLAB software.

Chapter-3 Methodology

3.1 Sentiment Analysis

The document-term matrix was formed for words in tweets and ranked based on high frequency. The highest frequency words appeared in the Wordcloud of Apple Inc.



3.2 Time-series Forecasting

The problem of forecasting the future price of securities on the stock market (or currency exchange rates, and so on). Markets have very different statistical characteristics similar to natural phenomena such as weather patterns. Machine learning and Deep learning combination can be used to forecast markets, through access to publicly available data and when it comes to markets, past performance can be used as a good predictor of future returns by using the changes in small seasonal intervals.

*3.2.1 ARIMA Model*

Machine learning is applicable to datasets where the past is a good predictor of the future by dividing past years data into small seasonal intervals like ARIMA model's moving distributed lagged dataset. ARIMA stands for a combination of Autoregressive Models (AR), Integrated Models (I), and Moving Average Models (MA) & Seasonal Regression Models. ARIMA is used in financial time series because it can be viewed as piecewise stationary or short-time stationary movements. It is a type of the Distributed Lags Model. The three models combined were:

*3.2.1.1 AutoRegression (AR) Models– A type of random process is represented by describing certain time-varying processes. The output variable of given time series is regressed on its own lagged values [Fig. 5.]. The number of time lags is denoted by the “p” value in the model.*

*3.2.1.2 Differencing or Integrated (I) Models – It indicates that the data values were replaced with the difference between their values and the previous ones, known as differencing. Distributed regression of the time series is involved to convert a non-stationary time series to a stationary one. The degree of differencing is denoted by the “d” value in the model.*

Non-seasonal ARIMA models are represented by ARIMA(p,d,q). Seasonal ARIMA models have another factor “m”, where m is the number of periods in each season. Some of the standard values for ARIMA are (1,1,0), (1,2,1), (1,0,0), (0,1,0). To get most accurate values for (p,d,q), the value on each of the small subparts of the time series curve was calculated and used for that part itself.

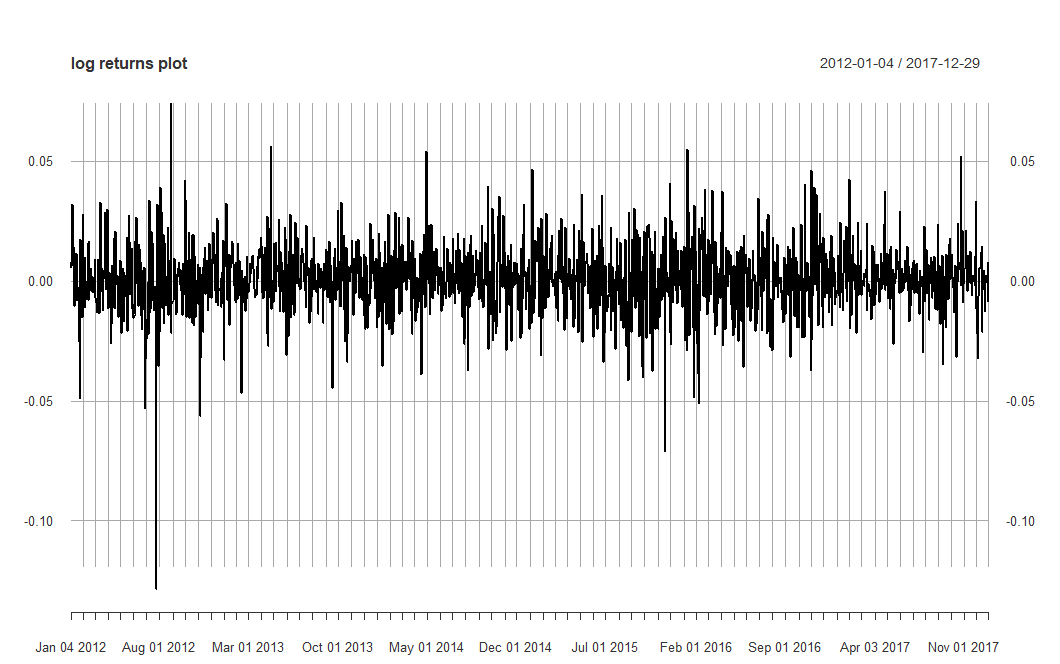
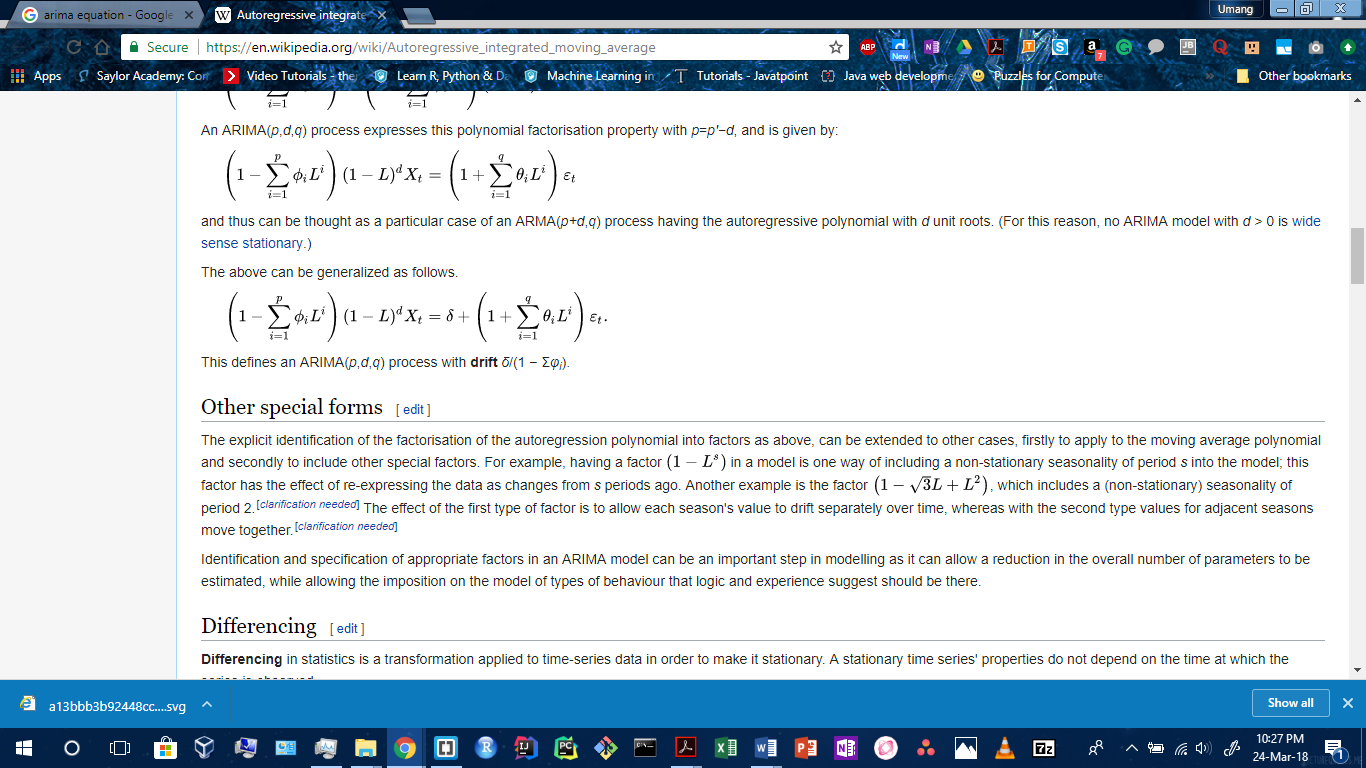


Fig.5. Logarithmic Returns Plot for Apple Inc. from Yahoo! Finance

*3.2.1.3 Moving Average (MA) Models – It gives regression error as a linear combination of past error terms. The number of lagged values of the error term is denoted by the “q” value in the model.*

The general equation for ARIMA can be given as-



where this defines an ARIMA(p,d,q) process with drift equal to δ/(1 − Σφi).

Here, ACF plot and PACF plot [Fig. 6.] were used to give the accurate lagged values of the error term for getting “q” value. The values of “p”, “d” and “q” were adjusted automatically in each seasoned interval, for improving the accuracy of the overall prediction. The testing of ARIMA gave idea of how well the model fits the data through AIC , AICc and BIC values. All the values obtained were very low, and lower the values of these terms, the better the ARIMA model fits. The obtained values of (p,d,q) were (2,1,2) respectively.

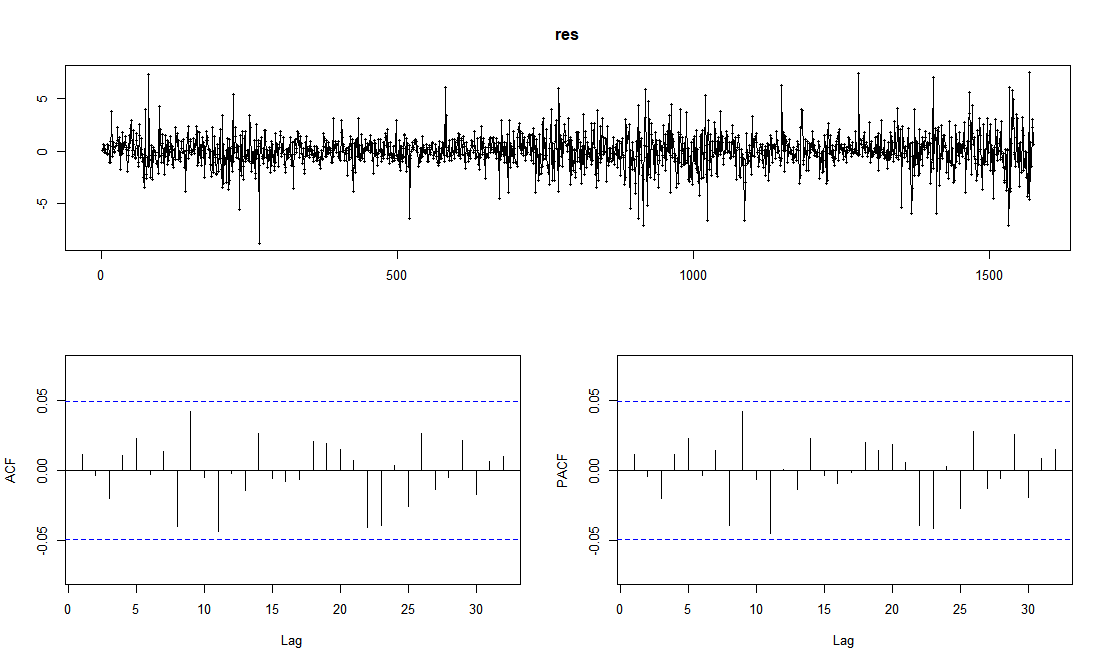
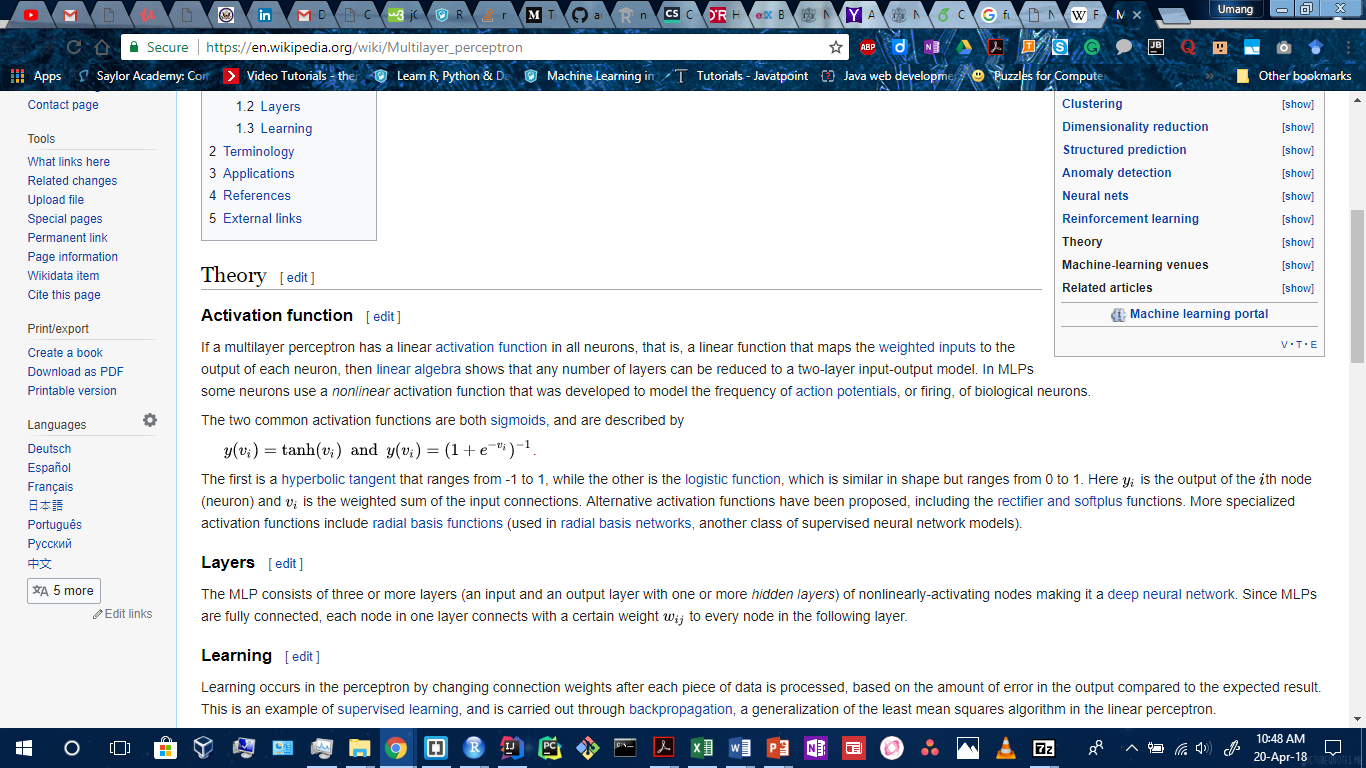


Fig.6. Residuals from Logarithmic Returns along with ACF and PACF Plots of them

AR(2) meant that logarithmic regression was based on previous 2 values, d=1 meant that Random Walk was present in the time series of Closing prices, and MA(2) implied that 2 regression values were used as a combination of lagged error terms in differencing. After that, an extensible time series (xts) object is initialized for Actual log returns and a data frame for the forecasted return series. A loop was then run through each seasoned interval for training and later testing the forecasted values of daily returns in logarithms.

*3.2.2 Neural Network*

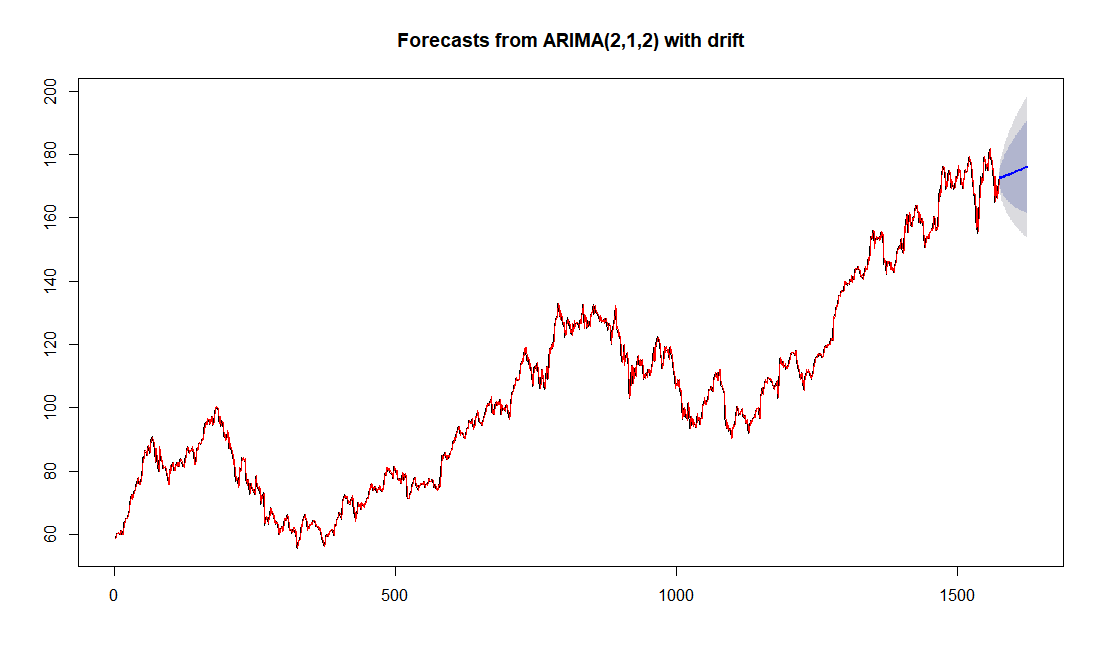
A feed-forward Neural Network was used on the forecasted logarithmic returns obtained from ARIMA through the (p,d,q) parameters passed to the Neural Network. The network was constructed with single hidden layer with ARIMA’s trained lagged inputs for forecasting and number of states equal to half of the stock prices for training. Number of non-seasonal lags passed from ARIMA results of ‘p’, and 20 networks were fitted with random start weights using logistic sigmoid function.



Equation of the logistic sigmoid function ranges from 0 to 1. Here y(v[i]) is the output of the ith node (neuron) and v[i] is the weighted sum of the input connections inside hidden layer, each node in one layer connects with a certain weight w[i,j] to every node in the following layer.

The model predicts the Closing price for next 50 days along with the logarithmic Daily change for any number of days. The results obtained predicted the market direction which was verified on the NASDAQ and forecasted Closing Price, which was very near to the actual value.

*3.2.3 Random Walk Model*



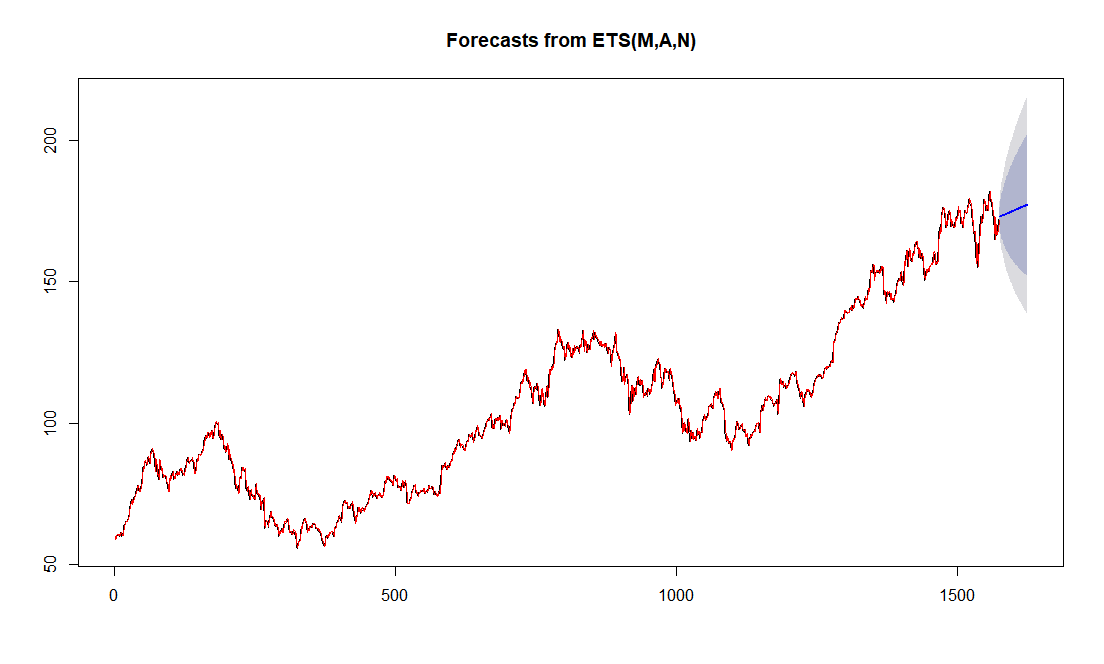
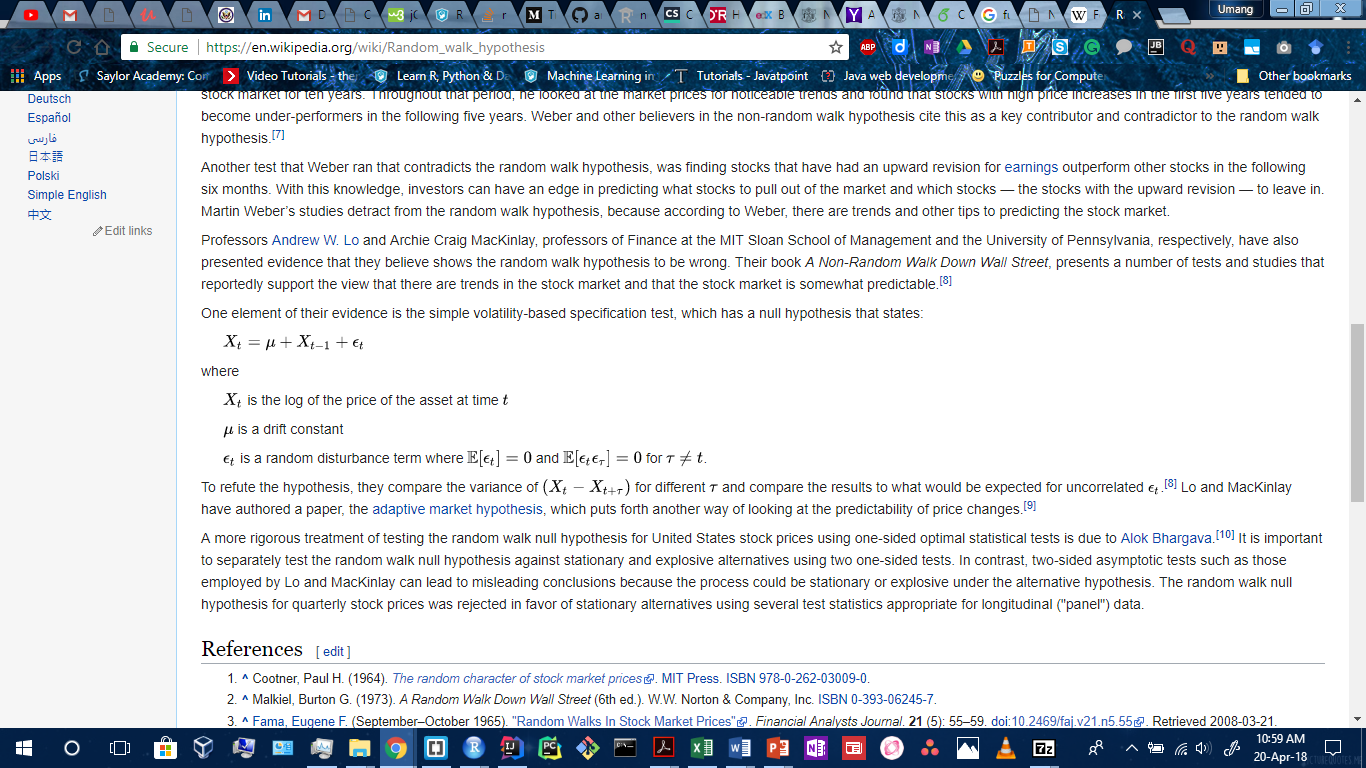


Fig.11. The Drift boundaries of ARIMA, verified by Exponential Smoothening (ETS) predicted by Random Walk of Ensemble Model

The Random Walk Model is used in Ensemble for finding the Drift boundaries obtained [Fig. 11.] in ARIMA forecasts. It is a famous theory among financial hypotheses which says that time-series data is formed from the random or stochastic process, that consists of a series of random steps on the financial historical dataset space such as the Daily Changes. It connects the next day’s opening price to be previous day’s closing price with most likely deviation to be maximum frequency/probability of percentage change which occurred on daily basis.



X[t] is the prediction of Deviation in current day, X[t-1] is previous day’s closing price, u is the most probabilistic frequent boundary value of drift and error term is added along with it for stationary time series.

Through a MapReduce implementation of frequencies of past Daily changes on Hadoop framework, the most likely boundary values of percentage change for the region of 95 percent and 90 percent occurring daily is found from the Random Walk model from the historical data of past 40 years.

Chapter-4 Results and Discussions

4.1. Sentiment Analysis Results

4.2 Time Series Forecasting Results from Ensemble Model

4.2.1 ARIMA forecasts

ARIMA used 97% data for training and 3% for testing from 6 years data. The Accuracy Percentage of the forecast was calculated by aggregating values in a table. The predicted stock returns for testing logarithmic returns data [Fig. 8.] are also compared with their actual returns in the testing phase of ARIMA. The accuracy was 50 percent until here and was increased through the usage of Neural Network in the Ensemble Model.

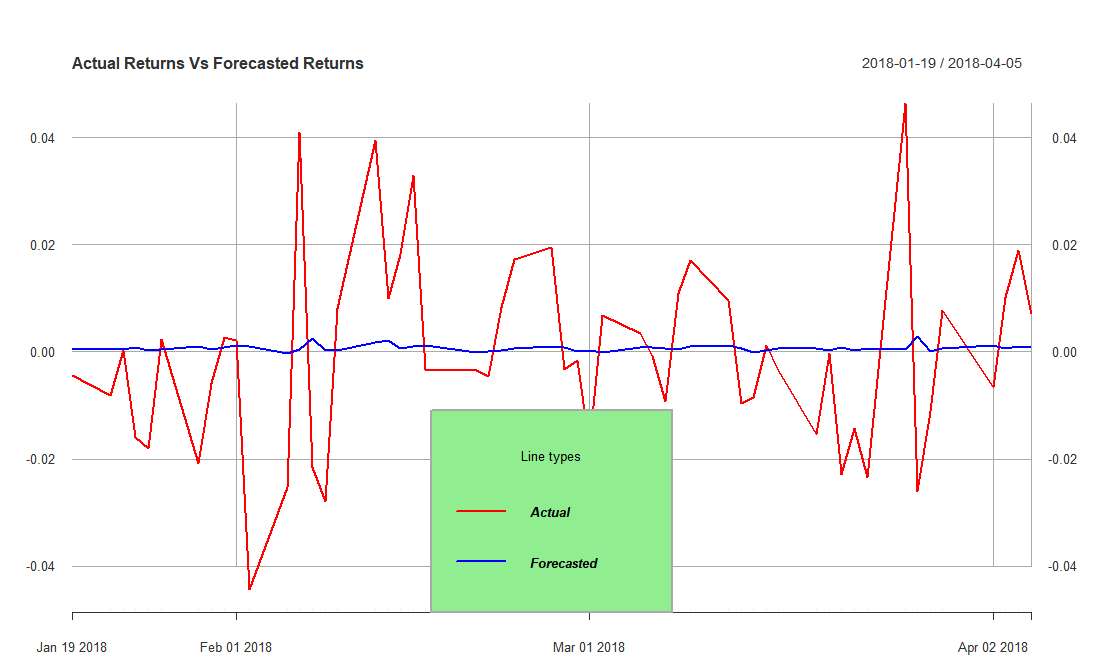


Fig.8. Actual Returns Vs Forecasted Returns for Apple Inc. from ARIMA

4.2.2 Neural Network forecasts and predictions

The testing proved the results to be more accurate from the Ensemble of ARIMA and Neural Network to be 89 percent accurate [Fig. 9.] with very low MAPE (Mean Average Percentage Error) = 2.75 and RMSE (Root Mean Square Error) = 5.43.

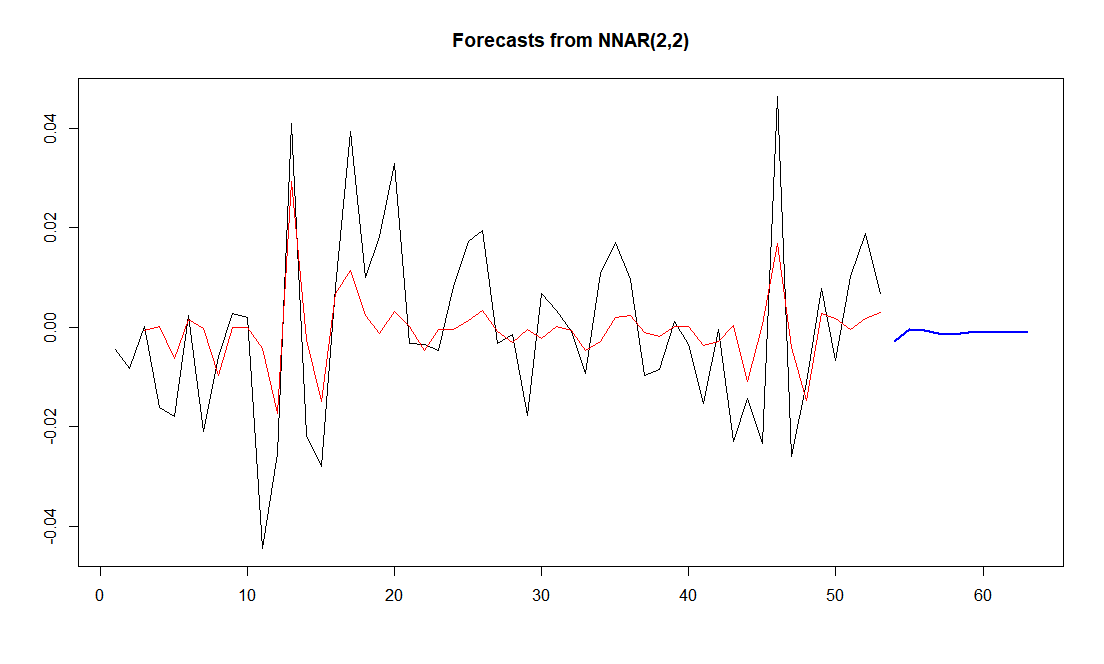


Fig.9. Actual Returns Vs Forecasted Returns for Apple Inc. from Ensemble of Neural Network and ARIMA

The model predicted the Closing price for next 50 days along with the logarithmic Daily change for any number of days. The results obtained predicted the market direction to gradually decrease which was verified on the NASDAQ and forecasted Closing Price for 9th April was 172.01 which was very close to the actual value of 170.05 [Fig. 10.] which was very near to the actual value. The drift in the forecast of 9th April is of magnitude 1.96. These drift boundary values with probabilities were calculated through Random Walk.

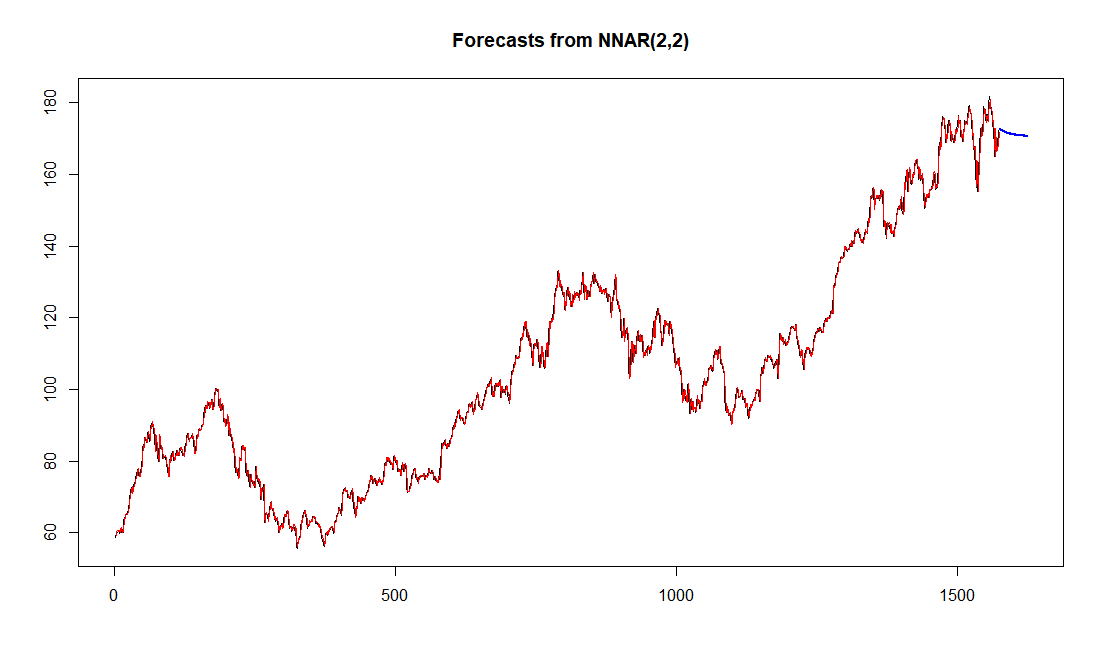


Fig.10. Closing Price of Next 50 days for Apple Inc. forecasted from the Ensemble Model

4.2.3. Random Walk Drift predictions

The drift boundary values were found from past 40 years data through Map-Reduce on Hadoop of Daily Change. The boundary values indicated a deviation of 10.11% to be most frequent for 9th April and the magnitude of 1.96 is well within 10.11% of 170.05 value of closing price.

So, the forecasted result from the Ensemble Model was 172.01 with 10.11% deviation for Apple Inc.’s Closing Price of 9th April.

Chapter-5 Summary and Conclusions

In this proposed project, an attempt is made to develop a prediction and forecasting model for finding the future stock market movements and their values based on the sentiment analysis of opinions and emotions expressed on Twitter feeds for a healthcare industry, like Apple Inc., using opinion mining and using this correlation for time series analysis using historical stock market data, big data processing, and data preprocessing, and machine learning algorithms and techniques. The predicted outputs show the proposed model’s potential to forecast the stock market movements for the short-term analysis of future, like for the next day, helping investors in their profitable investments in securities of stock markets and decisions related to buying/selling/holding a stock share, and thereby they can also contribute to advancements in technology through investing in the best Technology industry, and compete successfully with other emerging prediction and forecasting techniques. Other industry sectors like Healthcare could also be used for investments in saving lives.

5.1 Scope for FUTURE WORKS

Other opportunities for improvements, which could be done on the proposed model, include using sentiments from News sources. Also, outliers and robots can be identified by validating the people responsible for the sentiments from the past week to be humans or not. Speech recognition could also be combined in sentiment analyzer.

REFERENCES

[1] Rao, Tushar, and Saket Srivastava. “Analyzing stock market movements using twitter sentiment analysis.” In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pp. 119-123. IEEE Computer Society, August 2012.

[2] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of computational science 2, no. 1 (2011): 1-8.

[3] Makrehchi, Masoud, Sameena Shah, and Wenhui Liao. "Stock prediction using event-based sentiment analysis." In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, vol. 1, pp. 337-342. IEEE, December 2013.

[4] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1, no. 12 (2009).

[5] Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." In Proceedings of the 14th international conference on World Wide Web (WWW-2005), pp. 342-351. ACM, May 10-14, 2005.

[6] Sahni, Tapan, Chinmay Chandak, Naveen Reddy Chedeti, and Manish Singh. "Efficient Twitter sentiment classification using subjective distant supervision." In Communication Systems and Networks (COMSNETS), 2017 9th International Conference on, pp. 548-553. IEEE, June 2017.

[7] Angadi, Mahantesh C., and Amogh P. Kulkarni. "Time Series Data Analysis For Stock Market Prediction Using Data Mining Techniques With R." International Journal of Advanced Research in Computer Science 6, no. 6 (2015), August 2015.

[8] Chen, Ray, and Marius Lazer. "Sentiment analysis of twitter feeds for the prediction of stock market movement." stanford. edu. Retrieved January 25 (2013), Cs 229, pp. 15, 2013.

[9] Bing, Li, Keith CC Chan, and Carol Ou. "Public sentiment analysis in Twitter data for prediction of a company's stock price movements." In e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on, pp. 232-239. IEEE, 2014.

[10] Zhang, Linhao. "Sentiment analysis on Twitter with stock price and significant keyword correlation." Ph.D. dissertation, pp. 130, 2013.

[11] Dickinson, Brian, and Wei Hu. "Sentiment analysis of investor opinions on twitter." Social Networking 4, no. 03 (2015): 62..

[12] Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Standford University, CS229 (2011 http://cs229.stanford. edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSentimentAnalysis. pdf ) 15 (2012).

[13] List of Positive Words - Positive Words Research, http://positivewordsresearch.com/list-of-positive-words/.

[17] Bouktif, Salah, and Mamoun Adel Awad. "Ant colony based approach to predict stock market movement from mood collected on Twitter." In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, pp. 837-845. IEEE, 2013.

[18] Gunduz, Hakan, Zehra Cataltepe, and Yusuf Yaslan. "Stock market direction prediction using deep neural networks." In Signal Processing and Communications Applications Conference (SIU), 2017 25th, pp. 1-4. IEEE, 2017.

[19] Bautu, Elena, Andrei Bautu, and Henri Luchian. "Evolving gene expression programming classifiers for ensemble prediction of movements on the stock market." In Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on, pp. 108-115. IEEE, 2010.

[20] Nelson, David MQ, Adriano CM Pereira, and Renato A. de Oliveira. "Stock market's price movement prediction with LSTM neural networks." In Neural Networks (IJCNN), 2017 International Joint Conference on, pp. 1419-1426. IEEE, 2017.

Historical Data--

[14] NASDAQ Stock Quote for Apple Inc., https://www.nasdaq.com/symbol/aapl.

[15] Google Finance Apple Inc.'s NYSE Quote, http://www.google.com/finance?q=NYSE:AAPL .

[16] Yahoo! Finance Stock Details for Apple Inc., https://finance.yahoo.com/q?s=AAPL.