Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2018, 5(3): 162-169

Research Article ISSN: 2394 - 658X

Forecasting Stock Prices Using a Hybrid Approach

RMCDK Rajasinghe, WDNM Weerapperuma, WUNN Wijesinghe, KKKP Rathnayake and L Seneviratne

Department of Electrical and Computer Engineering,
Sri Lanka Institute of Information Technology (SLIIT), Colombo, Sri Lanka
lasantha.s@sliit.lk

ABSTRACT

Stock Market provides the basis for transactions between large business organizations and individual investors. Companies issue stocks to general public to raise funds while investors buy the stocks to gain profits. The Random Walk Hypothesis governs the Stock Prices as it changes constantly due to various factors. The research, Forecasting Stock Prices Using a Hybrid Methodology is carried out to implement a decision support system that provides insight for selecting profitable stocks using multiple forecasting mechanisms.

Keywords: Stock Market, Random Walk Hypothesis, Forecasting Algorithms

INTRODUCTION

Stock market is the financial platform that enables large organizations to sell portions of their company as shares in order to fund business endeavours [1-2]. Upon investing in a company, the investor becomes a shareholder and in return inherits a part of the companies' income. Shareholders are entitled dividends for the invested amount once the company generate a profit. Investing profitably is a tricky process where it requires an investor to be well aware of the company's economical background as well as the performance of the company. Investor should buy the stock before the price goes up and also be wise enough to sell the stock before the price fall back down. Fundamental factors such as the industry performance, Sentimental Factors such as the social opinion of the company and also the Economic Factors like Political Shocks and Inflation play a huge role in keeping the market efficient and non-stationary. So it is very hard to determine how companies will perform.

Shares are traded either through exchanges or over-the-counter-markets. In stock exchanges companies can issue stocks as well as the investors and shareholders can trade their stocks. It's a common platform for transactions between companies and investors as well as transactions among investors, more simply the exchange of securities is done in two levels, in the primary market stocks are first offered to investors and in the secondary market shareholders trade the stocks they possess among other investors [1-2]. So in investor's perspective, since they cannot control how they will gain profits from the companies they own, they can assume which company is likely to be more successful or which company is likely to fail and trade their stocks in the secondary market. This is where the three timing decisions come to play where the investor will have to make a judgment call on his investments; the three decisions are about determining the ideal time to sell a stock, ideal time to buy a stock and the ideal time to hold a stock for future profits.

Shareholders have to undertake the risk of losing the investment in making these three decisions correctly. Return on the investment will only be profitable if the performance of the company in the quarter is profitable. So, evidently, investors shoulder a huge financial risk in becoming shareholders, where the credibility of companies is very unpredictable and uncertain due to the dynamically changing nature of the world economies. The trader should be wise and experienced enough to gain profits by trading the stocks at the ideal time. Taking all the factors in to consideration one could arrive at a conclusion that investing in the share market either direct or indirect involves a lot of risk. The investor has to assume and accumulate the worth of undertaking the risk in order to earn profits. Currently there are many tools available that help the investors in answering the three questions more effectively.

RESEARCH METHODOLOGY

The forecasting process is designed in a way that it addresses and analyses most areas that can affect stock prices and the economics. Fig. 1 depicts the basic solution outline suggest by the research.

Generation of predictions are implemented under four main methodologies which include;

- Data Mining
- Statistical Analysis
- Graphical Analysis
- Sentiment Analysis

All the four methodologies will generate predictions that will be used in generating the final hybrid prediction from the Dynamic Weight Distributor. Final outcome from the process contains a result enriched with a wide range of aspects that moves the stock prices and market economies.

The main inputs for the process were historical stock quotes and twitter feeds regarding stock market.

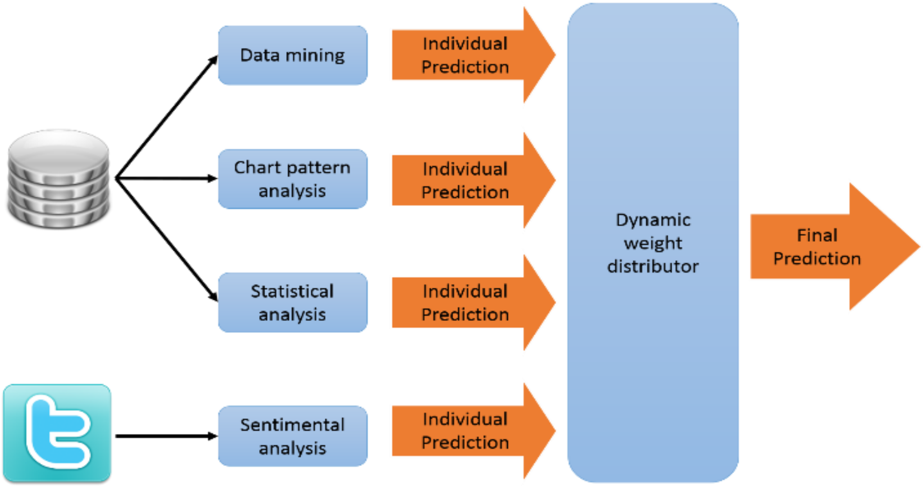


Fig. 1 Hi-Level Design of the Research

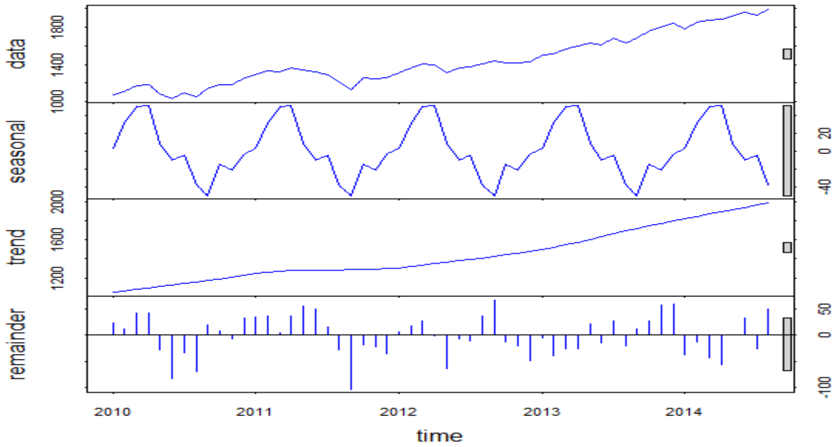


Fig. 2 Decomposed Time Series in to Data, Seasonal, Trend and Remainder

Data Mining

Stock Markets deal with multitude of data in a day and it is impossible to keep track of these data and identify patterns manually. Data mining can be used to extract the knowledge embedded within this data to provide a predictive analysis which will provide an insight to the investor that will invoke his judgment and experience in to the equation. Since stock market data is of time series data type it contains three main components: the trend, seasonal pattern and the irregular or remainder component. Loess data smoothing methodologies are used on data to further clarify the underlying patterns that are distorted due to seasonal and irregular elements. Once the data is decomposed using the seasonal-trend decomposition a seasonally adjusted set of data can be obtained using an additive model. Fig. 2 shows the three components in the time series data decomposed using seasonal-trend decomposition. The seasonally adjusted data is used to generate predictions using Naïve and Random Walk Hypothesis with Drift methods.

Naïve methods consider the most recent fluctuations in the data to determine price variation in the future using the seasonally adjusted data. The patterns in the recent data are projected to continue in the future. Fig. 3 depicts the output from the naïve forecast.

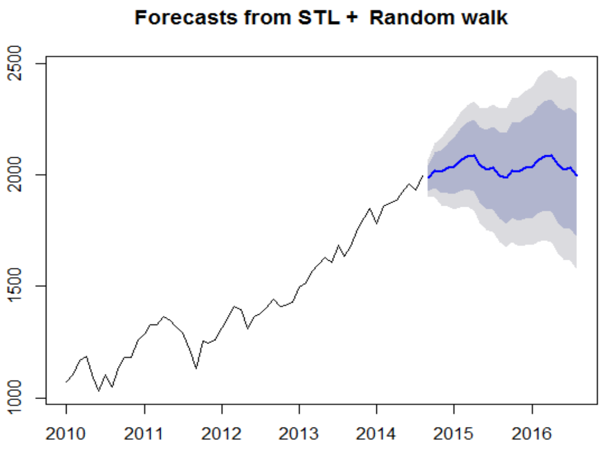


Fig. 3 Naive Forecast

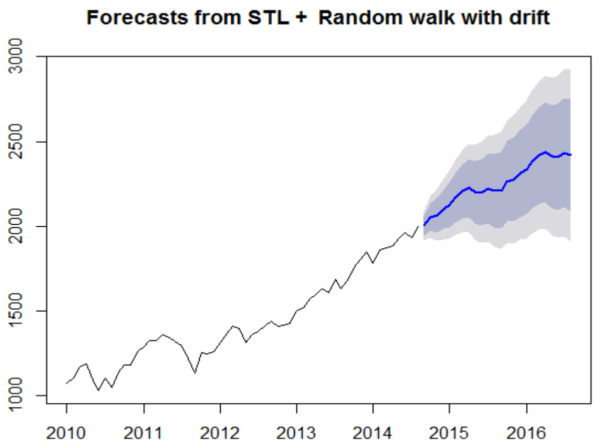


Fig. 4 Random Walk with Drift Forecast

Random walk with the drift methodology also uses the processed recent data but it also incorporates the drift of the trend to project the possible vales in the future. Fig. 4 depicts the Random Walk with Drift forecast output. These two methodologies provide a good insight on how the data will behave in the future by analysing the underlying patterns in data.

Statistical Analysis

The time series data is further analyzed in the Statistical Analysis process [3-4] using time series analysis methodologies like ARIMA. ARIMA or Autoregressive Integration Moving Average is a popular statistical methodology which aims to describe the autocorrelations in a dataset. A stationary time series is one whose properties do not depend on the time at which the series is observed. In ARIMA (p,d,q), p is the number of autoregressive terms, d is the number of non-seasonal differences, and q is the number of lagged forecast errors in the prediction equation. ARIMA models are defined for stationary time series and if the data is not stationary, differencing process is done until stationary time series is obtained. If the differencing process is done d times, then in ARIMA (p,d,q) model, d will become the order if differencing used.

In a multiple regression mode, the variable of interest for the forecasting is combination of predictors. In an auto-regressive model, forecast the variable of interest using a linear combination of past values of the variable. Term auto-regression indicates that it is a regression of the variable against itself. In ARIMA, p can be written as,

$$y_t = c + \phi_1 y_t - 1 + \phi_2 y_t - 2 + \cdots + \phi_p y_t - p + e_t, \quad (1)$$

Here *c* is a constant and e_t is white noise. This is similar to multiple regression but with *lagged values* of y_t as predictors. This is referred to as an AR(*p*) model**.** Rather than using past values of the forecast variable in a regression method, a moving average model uses past forecast errors in a regression-like model. In ARIMA, q can be written as,

$$y_t = c + e_t + \theta_1 e_t - 1 + \theta_2 e_t - 2 + \cdots + \theta_q e_t - q, \quad (2)$$

Where e_t is white noise. This is referred to as an MA(*q*) model. So the ARIMA model would be, [5]

$$y'_t = c + \phi_1 y'_t - 1 + \cdots + \phi_p y'_t - p + \theta_1 e_t - 1 + \cdots + \theta_q e_t - q + e_t, \quad (3)$$

The auto.arima() methodology is used in the implementation and it is proved to be very useful, but anything automated can be a little risky, and worth understanding the behavior of it. The auto regressive model will choose a model for the forecasting. In above equation, constant C has an important effect on the long term forecasts,

- If *c*=0 and *d*=0, forecasts will go to zero.
- If *c*=0 and *d*=1, forecasts will go to a non-zero constant.
- If *c*=0 and *d*=2, forecasts will follow a straight line.
- If *c*≠0 and *d*=0, forecasts will go to the mean of the data.
- If *c*≠0 and *d*=1, forecasts will follow a straight line.
- If *c*≠0 and *d*=2, forecasts will follow a quadratic trend.

The value of *d* also has an effect on the prediction intervals, the higher the value of *d*, the more rapidly the prediction intervals increase in size. The value of *p* is important if the data show cycles.

Graphical Analysis

The Graphical Analysis is a part of Technical Analysis and it can be used to identify stock patterns in past datasets. Feature selection is a pre-processing step that aims to select the most relevant subset of attribute by eliminating unrepresentative attributes from the dataset. Stock data patterns are the best way to represent trends in a stock. Even someone with less knowledge in stock price variations can analyse chart patterns without many troubles. These patterns are further classified in to two categories. Those are reversal and continuation patterns. Each pattern which, occurred in past may have a certain similarity to present. The goal behind this analysis is finding an appropriate relationship between the past and current patterns. This analysis is done by using three main sub-systems. Those are,

- Candlestick pattern analysis
- Algorithmic analysis for stock pattern recognition
- Artificial Neural Network based pattern recognition

Candlesticks is a graphing methodology which represents data in a more descriptive manner. A sample candlestick chart for Stocks of Apple Organization is depicted in Fig. 5. A candlestick is based on four variables. They are open, high, low and closing prices of a stock. It's further categorized into increasing or decreasing. If the closing price of a stock is higher than the opening one, it implies increasing candlestick. The decreasing is the vice versa of increasing. There are several types of standard patterns found in candlestick charts such as doji, gravestone doji, dragonfly doji, hammer patterns and long shadows. Based on these patterns candlestick pattern analysis system is working as supporting system, which is responsible for determining whether a stock trend will reverse after a special pattern such as doji occurs. This enables the decision making process of the dynamic wright distributor much simpler. Below are the steps of the process,

- Analysing past stock data and retrieve dataset containing all the doji occurrences
- Consider all the unacceptable data points one by one and check data points of days following the pattern and before the pattern and determine whether the underneath trend reversed
- Get same calculations for all patterns and develop a constant value which is dedicated for a stock
- Whenever the relevant stock is used for the prediction, relevant calculated constant can be used for validate the generated prediction

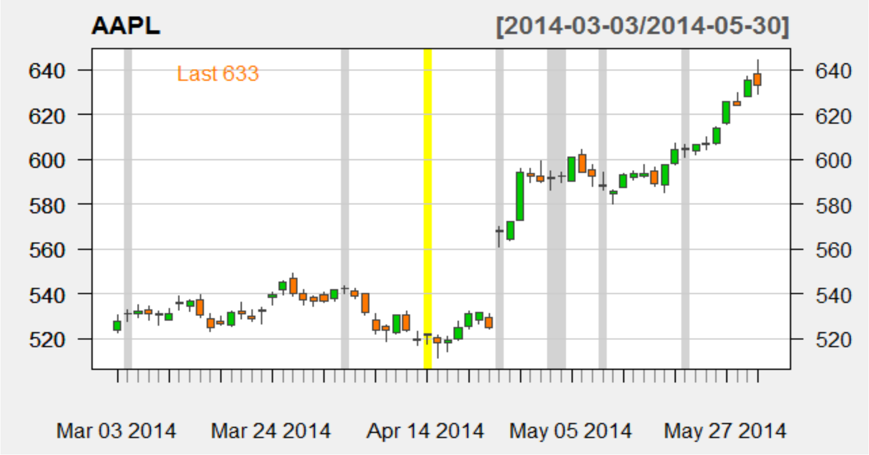
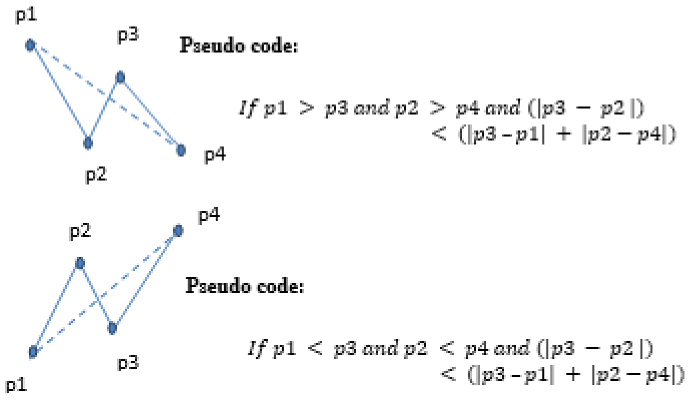


Fig. 5 Candlestick patterns in Apple time series

The stock data considered as noisy and rapidly varying. In order to identify patterns using the algorithmic graphical analysis, the first step is to filter out the noise. Sliding window mechanism is used for filtering. This technique is based on turning points that change the trend of a stock dataset. It's done in two levels. First row stock data will be considered and fragmentation of the dataset into smaller parts is done. Then consider a single part and identify turning points by using local maximums and local minimums. In level one, for each segment only local minimums and maximums are taken and points which does not affect for the under laying trend are ignored [5]. After selecting the points in separate segments, merging process takes over, combine neighbouring segments one another, and generate a new dataset. In level two, the dataset from level one is used. This extract points which are more contribute for the under laying trend. This is done by using an algorithm. Below is basic step for eliminating intermediate points. P1, P2, P3 and P4 are consecutive data points in the stock dataset.

After accruing filtered dataset, pattern recognition algorithm works on it. The algorithms are able to extract head and shoulder patterns, inverse head and shoulder patterns, broadening patterns like flat-top broadening, flat-bottom broadening and semantic broadening and triangle patterns like, flat-top triangles, flat-bottom triangles and semantic triangles. In each of above mentioned patterns, formed using seven consecutive data points. Each seven data points for each pattern has its own characteristics. The algorithms use those characteristics for the identification of a pattern [6]. Fig. 6 and Fig. 7 illustrate how the algorithm evaluates the data and identify a head and shoulder pattern. The threshold values can change accordingly and those will ensure how clear a pattern going is to be. The algorithm will check segments, which have seven segments iteratively and identify patterns in them. Prediction will be calculated using the identified patterns in the historical dataset by analysing next data points after a certain pattern. The calculation is done to form a price variation percentage for a certain stock from the end of a pattern.



Pseudo code:

```
If $p1 > p3$ and $p2 > p4$ and $(|p3 - p2|) < (|p3 - p1| + |p2 - p4|)$
```

Pseudo code:

```
If $p1 < p3$ and $p2 < p4$ and $(|p3 - p2|) < (|p3 - p1| + |p2 - p4|)$
```

Fig. *6* Filtering mechanisms

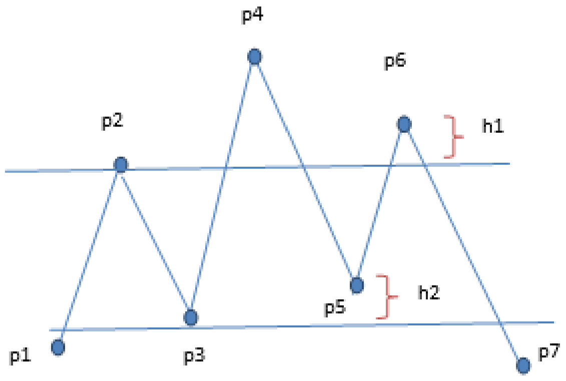


Fig. 7 Point identification in a head and shoulder pattern

Pseudo code:

```
Get all min points {p1,p3,p5,p7}
Get all max points {p2,p4,p6}
If h1 <  TRESHOLD_HEIGHT and
h2 < TRESHOLD_HEIGHT and
p2 < p4 and p6 <  p4
Then,
The segment is valid pattern
```

The neural network in this solution is capable of finding stock price patterns as the algorithmic solution but more efficiently [7-8]. Neural network has three layers. Input layer, which has seven neurons each accepting a stock value of series. Middle layer contains two hundred neurons and output layer has three neurons, each dedicated for a specific pattern. The design is a feed forward neural network, which has Sigmoid as the learning algorithm [9]. Dataset for learning is generated and taken from a data generation algorithm, which will produce appropriate sets of data, which suit specific pattern.

The steps taken for identifying a pattern from a dataset [10],

- Filtering the data
- Segmentation
- Inserting each segment to the neural net to identify patterns

Because the neural net is pre trended, no dynamic training involved [11]. So it will reduce the over training problem. Fig. 8 is a sample pattern identified by the neural net for apple stock, which shaped from 18th of Oct 2000 to 20th of Nov 2000. Prediction methodology is same as the one in algorithmic method [12].

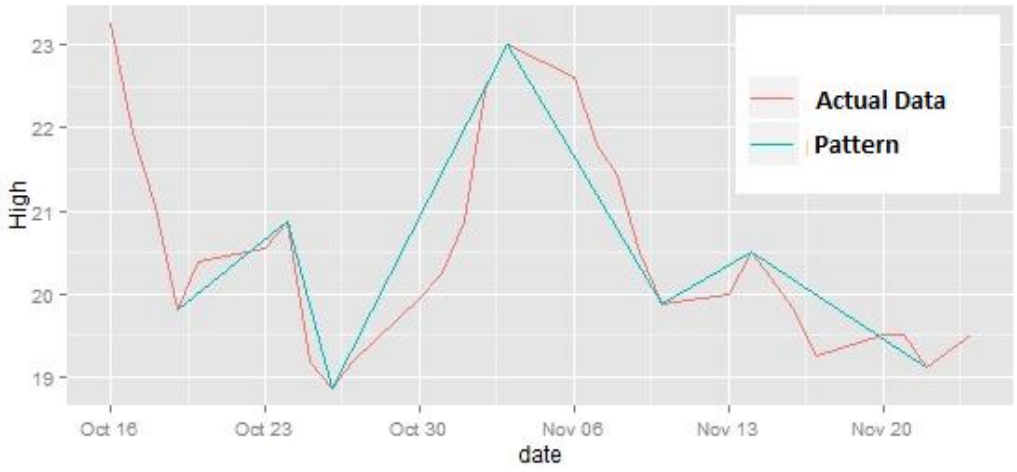


Fig. 8 Patterns Identified by NN

Sentiment Analysis

Human Emotions play a huge role in stock market as the investors are often manipulated with new information and trends in the market. In Sentimental Analysis approach the reaction of the society toward an action of a company is inspected to determine the psychological effect it has on the investors to invest more or invest less in a company. Extracting the public opinion is a tricky process as emotions and opinion is subjected to change from one person to another so for implementing this practically, Twitter website is used as it is the largest microblogging website in the world. Twitter data is obtained from twitter API based on a certain scenario or a company. The data is then processed and mined to extract the polarity and the emotion that is embedded within it. This is done using a data dictionary that classifies positive and negative words and world combinations. The obtained data will be mapped with the dictionary and the percentage of polarity is calculated to represent public acceptance or denial to the scenario.

Hybrid Weight Distribution

The final step of the research methodology is to merge all the results and generate a single forecast for a given stock. This is the responsibility of the weight distributor to allocate appropriate weights to each prediction and generate a single prediction which the investor can rely on easily. The weight distributor is implemented in a way that it takes the accuracy of the predictions in to consideration when allocating the weights. Accuracies of different approaches may change from company to company as well as from season to season. As there are many factors that affect the stock prices, the weight distributor finds the impact these factors have on the prices in different contexts and scenarios. The observations of correlations between prices and these factors that affect the prices are used as indicators in generating the most accurate forecast through the weight distributor.

RESULTS AND DISCUSSION

The results of the two methods from Trend Smoothing tend to go close but when considering the results separately, Naïve based method can deliver predictions up to two years by projecting the last obtained actual value. The main problem with this methodology is that as the prediction length goes beyond three months, the accuracy of the prediction tends to decrease drastically. The Random Drift Method Performs better when considering the predictions for long duration. So these two methods are used in parallel to get a good understanding about how the market would react in future. Table 1 and Table 2 define how the predictions of the two modules vary within the next 3 months.

Table -1 Predictions from the Naïve Method

Month	Low Range	Mean Prediction	Higher Range
September	1909.13	1974.39	2039.64
October	1892.03	1987.31	2076.6
November	1874.62	1987.64	2100.66

Table -2 Predictions from the Random Drift Method

Month	Low Range	Mean Prediction	Higher Range
September	1909.13	1974.39	2039.64
October	1898.03	1984.31	2076.6
November	1882.86	1997.84	211.83

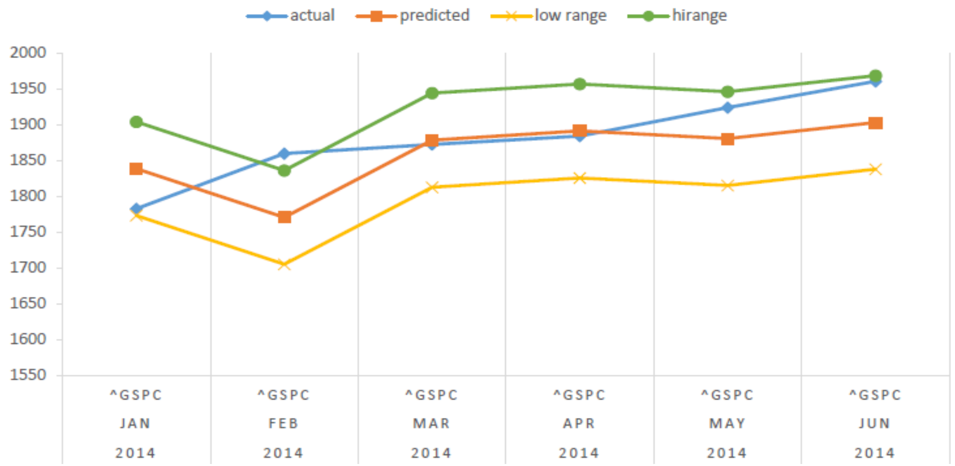


Fig. 9 Predictions and Observations - Monthly

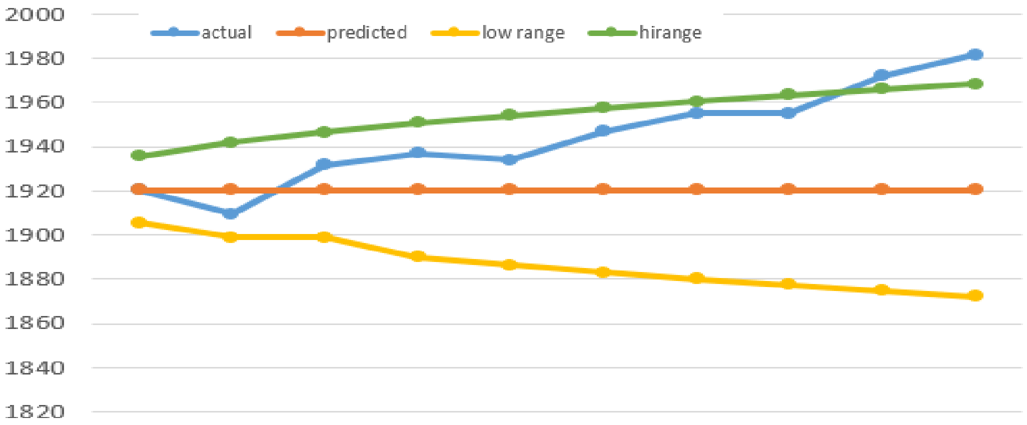


Fig. 10 ARIMA model prediction variance

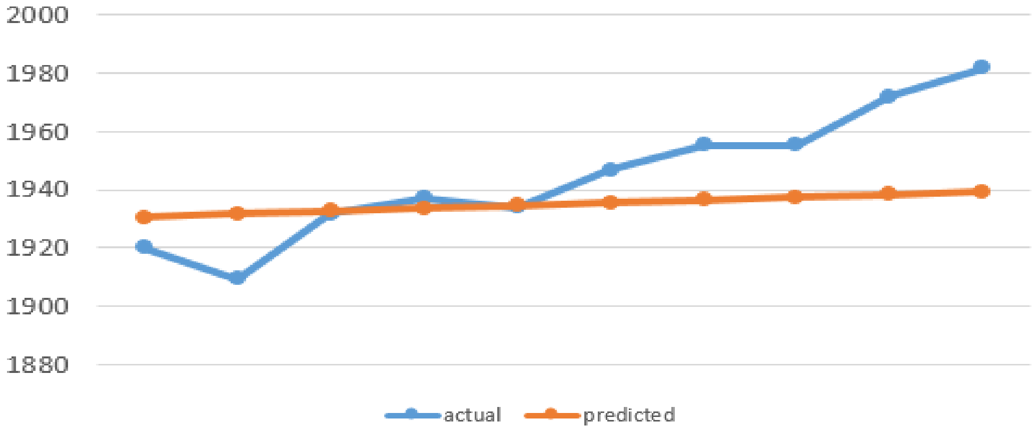


Fig. 11 GARCH model prediction variance

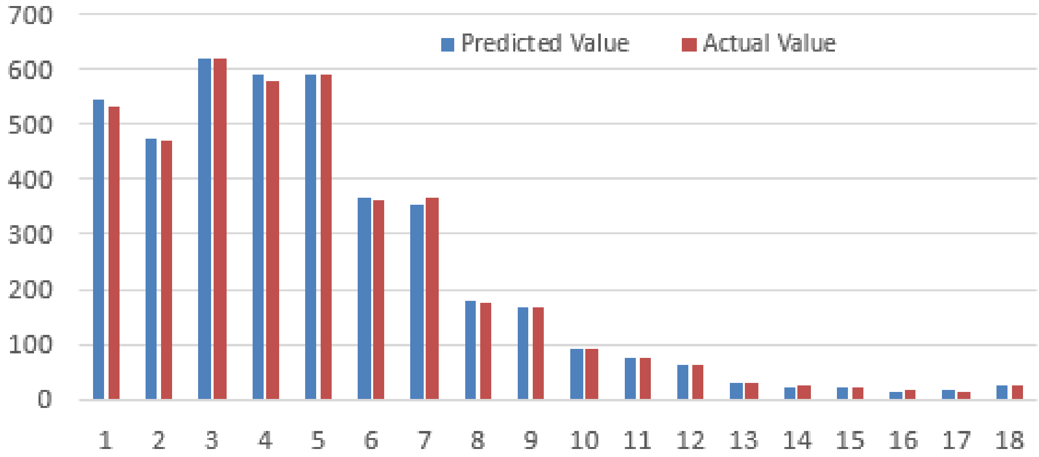


Fig. 12 Actual and predicted values from pattern recognition

Once the results are considered with the actual observations an idea about the accuracy of the results could be gained. The predictions and actual result variation is shown in Fig. 9. Fig. 10 depicts how the stock prices of S&P 500 have varied with the ARIMA model. The predicted daily data for 10 days are compared. Fig. 11 depicts how the S&P 500 index has behaved with regard to the predictions from the GARCH model. The predicted daily data for 10 days are compared. Chart pattern analysis only produce results when a certain pattern occurs in a dataset. Fig. 12 represents the results acquired from algorithmic pattern analysis on Apple dataset from 2000-Jan to 2014-Aug. The x-axis depicts number of patterns and y-axis depicts the prices. Since all these methods are inputted to a weight distributor methodology including the sentiment analysis, the final generated prediction will be enriched with various standpoints that would affect the stock prices in different proportions. The scalability of the application will not be a problem and the market for this kind of application is clearly evident.

CONCLUSIONS

The Research was based on the Random Walk Hypothesis which suggests the unpredictable nature of prices in financial market. The hypothesis was tested using four main modules in the research which included both purely quantitative and qualitative aspects that affect the stock price variations. The results obtained in the process suggest the possibility of the predicting the stock prices even though it is completely governed by the Random Walk Hypothesis. The results prove the prices can be predicted to some extent. A mechanism can be very effective which can be used to map other irregular factors that can affect the stock market other than public opinion itself. The same kind of mechanism can be used to find out how stock market has reacted to political, geographical and natural phenomena's. The Hurricane Katrina that happened in the mid of 2005 is a perfect example to this as it directly affected the world economies in the long run. The solution can be further developed to incorporate the irregular indirect factors along with the current features which mainly focus on the direct influences on the stock market.

REFERENCES

[1] Y Bayar, A Kaya and M Yıldırım, Effects of Stock Market Development on Economic Growth: Evidence from Turkey, *International Journal of Financial Research*, **2014**, 5, 93-100.

[2] B Ake and RW Ognaligui, Financial Stock Market and Economic Growth in Developing Countries: The Case of Douala Stock Exchange in Cameroon, *International Journal of Business and Management*, **2010**, 5, 82-88.

[3] JH Pedersen, *ARMA - GARCH Estimation and Forecast Using Rugarch, Web. http://www.unstarched.net/wp-content/uploads/2013/06/an-example-in-rugarch.pdf*, **2013**.

[4] J Yin, YW Si and Z Gong, Financial Time Series Segmentation Based On Turning Points, *International Conference on System Science and Engineering*, **2011**.

[5] D Banerjee, Forecasting of Indian Stock Market Using Time-Series ARIMA Model, *2nd IEEE International Conference on Business and Information Management (ICBIM)*, **2014**, 131-135.

[6] I KnowFirst, *I Know First: About Us*, Web. http://iknowfirst.com/I-Know-First-About, **2017**.

[7] W Andrew, H Mamaysky and J Wan, Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation, *Journal of Finance*, **2000,** 4, 1705-1765.

[8] N Srivastava, G Hinton, A Krizhevsky, I Sutskever and R Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Over fitting, *Journal of Machine Learning Research*, **2014**, 15, 1929–1958.

[9] MC Angadi and AP Kulkarni, Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R, *International Journal of Advanced Research in Computer Science*, **2015**, 6, 104-108.

[10] A Camara, W Feixing and L Xiuqin, Energy Consumption Forecasting Using Seasonal ARIMA with Artificial Neural Networks Models, *International Journal of Business and Management*, **2016**, 11, 231-243.

[11] S Benkachcha and J Benhra, Seasonal Time Series Forecasting Models based on Artificial Neural Network, *International Journal of Computer Applications*, **2015**, 116, 9-14.

[12] HH Myoung and M Byung-Ro, The Evolution of Neural Network-Based Chart Patterns: A Preliminary Study, *26th International Conference on Genetic Algorithms (GECCO)*, Berlin, Germany, **2017**.