STOCK PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Sumeet Sarode[1], Harsha G. Tolani[2], Prateek Kak[3], Lifna C S[4]

Computer Engineering Department, Vivekanand Education Society's Institute of Technology, Mumbai, India [1,2,3,4]

Abstract

In today's economy, there is a profound impact of the stock market or equity market. Prediction of stock prices is extremely complex, chaotic, and the presence of a dynamic environment makes it a great challenge. Behavioural finance suggests that decision-making process of investors is to a very great extent influenced by the emotions and sentiments in response to a particular news. Thus, to support the decisions of the investors, we have presented an approach combining two distinct fields for analysis of stock exchange. The system combines price prediction based on historical and real-time data along with news analysis. LSTM (Long Short-Term Memory) is used for predicting. It takes the latest trading information and analysis indicators as its input. For news analysis, only the relevant and live news is collected from a large set of business new. The filtered news is analyzed to predict sentiment around companies. The results of both analyses are integrated together to get a response which gives a recommendation for future increases.

Keywords: Stock Price Prediction, Stock Market Trends, LSTM (Long Short-Term Memory), Forecast of Stock Prices, Support Vector Machine (SVM), Efficient Market Hypothesis (EMH)

I. INTRODUCTION

For many decades, the object of studies has been the prediction of the stock markets and despite it's convolutedness, dynamism, and derrangeness, making it an extremely strenuous task. Various factors, huge amount of data, and a trivial signal to noise ratio are to be considered for an efficient prediction model which makes the task of predicting the behaviour of stock market prices significantly difficult. Nonetheless, there are a large variety of approaches presented in order to reach that goal.

Recurrent neural networks (RNNs) are used whenever the model asks for processing time series data or natural language. LSTM being one of the most successful RNNs architecture has the capability of giving distinct weights for every example, and inadvertently neglecting the memory that it considers to be irrelevant for predicting the next output so as to differentiate between currently operational and previous examples. It has proved very effective while handling such problems. So in contrast to other recurrent neural networks,it is more efficient in handling long input sequences. Thus, with the help of LSTM network, a very high level of accuracy can be achieved in predicting the upcoming trends and the price estimates of various stocks.

All sorts of information can be accessed in this closely connected world making it not very difficult to make informed choices about any particular topic. The type of news that we encounter on a daily basis affects most of the decisions that we make. To take a proper decision towards a particular entity the sentiment becomes the driving force. Major news significantly impacts the traders' investment and thus it calls for a major change in the investment plan. Thus, while predicting the stock prices with the help of historical data it becomes necessary to consider the sentiment analysis of the news on a regular basis.

II. LITERATURE SURVEY

A. News Analysis

Nowadays, the stock market trend prediction is being explored by many research groups using social media analytics.To get the sentiment of every good in news/tweet its polarity has been found. Polarity of any news/tweet can be calculated by the two available approaches- the dictionary-based approach, and the semi-supervised algorithm. In dictionary based algorithm, the polarity of every word in the news is calculated by comparing it with the words present in a predefined dictionary. The semi-supervised algorithm, as discussed by K. Mizumoto et al. in his paper[1], suggests that words be grouped as either positive or negative based on their occurrence in a manually built dictionary.

A similar research conducted by Robert P. Schumaker[2] on news articles and stock quotes from S&P 500 over a period of one month. Altogether, 9211 financial news articles and 10,259,042 stock quotes were gathered. These quotes and news were analyzed with respect to three textual representations, and the terms which occurred three or more times in an article were retained. The process of filtering was carried which caused the following breakdown:

—Bag of words used 4,296 terms from 2,839 articles
—Noun phrases used 5,283 terms from 2,849 articles
—Named entities used 2,856 terms from 2,620 articles

Using the Sequential Minimal Optimization method, this breakdown was processed by SVM derivative to achieve an accurate analysis of the news. Closeness and directional accuracy where the two chosen evaluation metrics.They have three different models. For making predictions, the first model(M1) was used. And within this model, no baseline stock price exists.The second one (M2) used the price of the stock at the very moment an article was published publicly. The third model (M3) used the extracted terms along with the +20 minute stock price regression estimate. For their prediction, all of the models mentioned above make use of the terms in the article.

B. Efficient Market Hypothesis

Efficient Market Hypothesis (EMH) (Fama 1964) is one amongst those various methodologies which aim to ascertain the upcoming trends in Stock Market. In EMH[2], it is assumed that all of the information available to everyone ascertains the price of a security. Based on the inputs taken, EMH is categorized into three forms- Weak, Semi-Strong, and Strong. Strong form of EMH takes into consideration all the historical trading details along with the public, and private information. The Semi-Strong form is impeded by incorporating past and present publicly announced information besides the price. The Weak EMH deals only with the price and the historical details.

C. Stock Price Forecasting using Machine Learning Techniques

Most research work was done on Artificial Neural Networks (ANN) with forecasting using machine learning[3]. In an ANN, many nodes are interconnected to simulate each neuron. The architecture of these networks consists of function specific layers such as the ones for input and output, and the processing layer. All the connections are assigned their own weights. The inputs of the ANN determine the output weights. While training the machine, patterns are identified and all the weights are changed. Kar in his paper[4] demonstrates the accuracy of the ANNs especially when there are no abrupt changes in the data. Patel and Yalamalle in their paper have agreed that an accuracy of above 50% can be achieved using ANNs.

III. METHODOLOGY

A. NEWS ANALYSIS:

Real-time news that is collected from various websites providing financial news is stored in a database. The collector in the analyzer is used for collecting the main body of the news from these websites. The database stores this data in XML format. The collected corpora contain many triggering words or phrases such as the name, symbol of the company, the issues related to the company. This is provided as input to the classifier. It is responsible for trimming the stored data. It trims the stored corpus in order to reduce it to relevant information only. The results from the classifier are provided to the Analyser. Here the sentiment analysis is performed using the dictionary-based approach on the parsed sentences. The Dictionary-based approach provides a very accurate result.Scores of the words present in the sentences are used to calculate the aggregate scores. The lexical analyzer tokenizes the news gathered to obtain an array of scores. The scores are calculated by matching each word with the dictionary. The sentiment score is the difference between the positive and the negative matches. After getting a score for each sentence, the sentences with the same score are put together with a specific score. Estimator is used for calculation of total score. The total positive/negative indicator is given by the product of the positive/negative score and the total number of sentences. The degree positivity, Negativity or Neutrality is obtained by the percent result of the respective indicator divided by the total score.

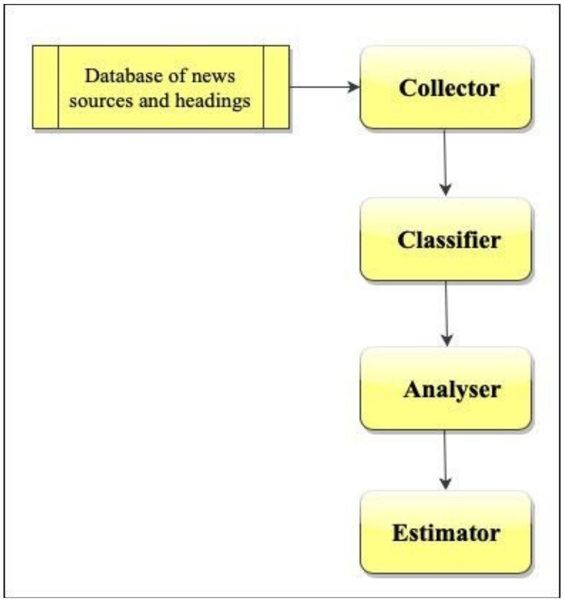

Fig. 1: Flowchart of our methodology

B. PREDICTION:

A sequential troupe of daily trading details of stocks over a set time period of N days is defined in the LSTM model. These daily details in sequence describe the trends of the stock with the attributes like the day high/low, open price, closing price and trade volume on a specific day within the N days. Comparing in sequence the closing prices of 3 consecutive trading sessions with that of the last day, the earning rates were calculated. The model comprises of two layers namely, an input layer which consists of number of cells equal to the sequence learning attributes that one sequence may hold, the LSTM layers, a compact layer, and an output layer consisting of similar number of cells. There are 4 types of learning features that could be given to the LSTM models. They are:

1. the historical trade details.
2. the technical analysis derived from these historical trade details.
3. the movement of the market indices.
4. the economic fundamentals.

First and the third type of data is essential for the forecasting of the prices of different stocks.

The LSTM model has an edge over the other models as it has improved memory cells. These cells are linked to each gateway. It also contains Forget gates. The connection to the memory stick is controlled by the forget gates. They are also responsible for remembering the error as per requirement and scale the feedback by each step[9]. It can also store information of any size. The input to the LSTM model is provided in the form of vectors. This vector comprises of elements such as the previous closing price, open price, high and low prices, the trade volumes of the current session. The vector also contains the results of the news analysis. The output layer of the model gives the final predicted price over the specific time periods.

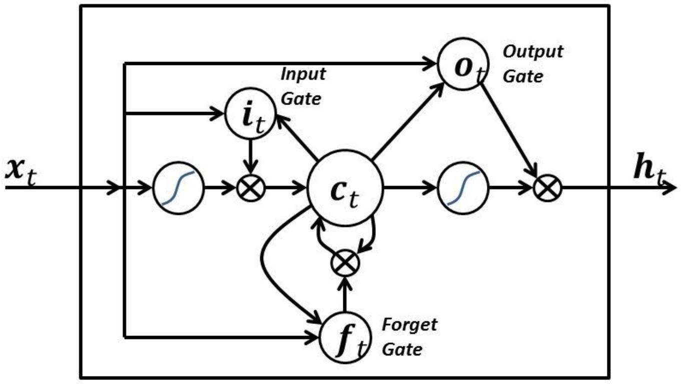

Fig. 2: Long Short-Term Memory Architecture[9].

IV. BLOCK DIAGRAM OF THE PROPOSED SYSTEM

Fig 3 refers to the block diagram of our proposed system.The real time news is collected with the help of News sources and is fed into the system.The News Analyzer Package of Python is being used to analyze the news and pass it for further processing. In the next step,Stemming is performed in which the words are simplified to its root known as lemma. The aim of this step is to get the keywords which can be extracted by the Keyword Extractor.

Further sentiment analysis is performed on the words where pieces of texts are classified as either positive or negative collecting the data from Domain Specific

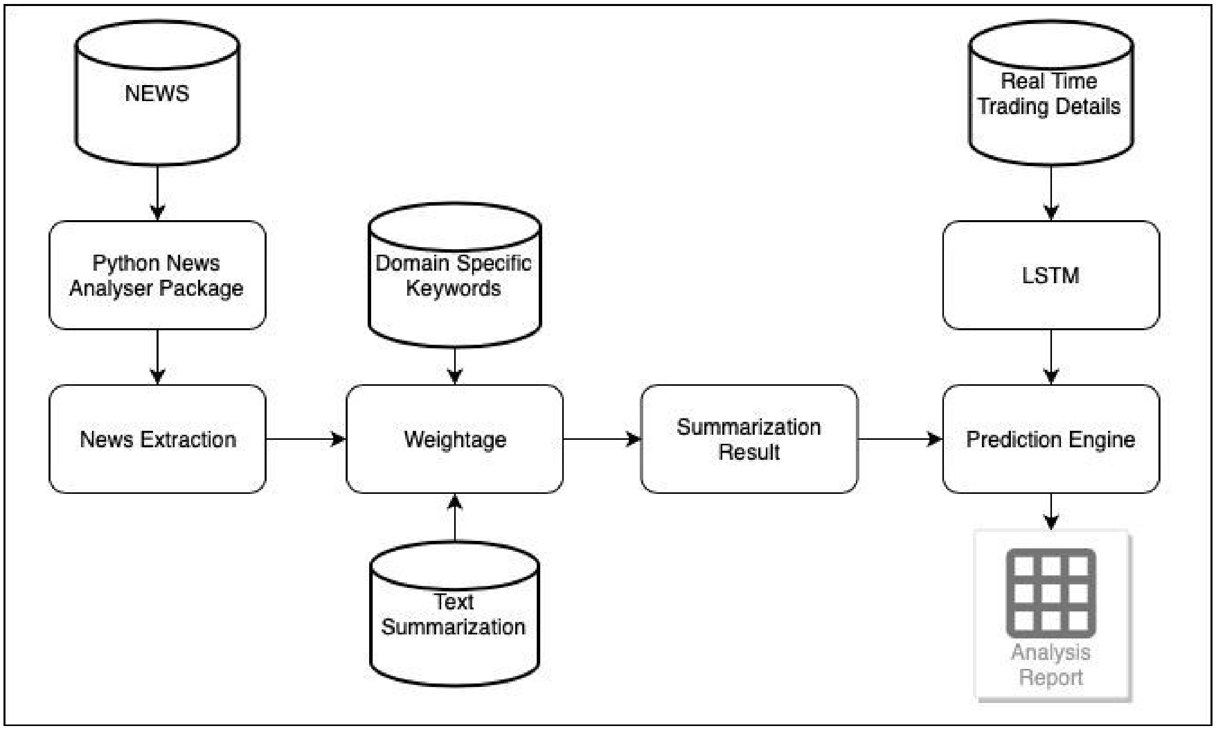

Fig 3: Block Diagram of Proposed System.

keywords database. Also their relation among themselves is also considered. The classified result is then stored in Text Summarization database along with the timestamp and ID for that particular news for further reference. And this whole process gives a summarized result based on the sentiments of the news.

On the other hand the historical and real time stock market prices are collected with the help of NSE(National Stock Exchange) Tools and further processed with the help of LSTM Model to predict the upcoming trends and prices. And the predicted results are then integrated with the summarized results of the news analytics with the help of different parameters and an analysis report is generated by the system specifying the names of the stocks and their change percent.

V. CONCLUSION

Thus, this paper proposes a system that would provide recommendations for the buying of shares of distinct companies. A decision making algorithm is prepared by using artificial neural networks and also taking into consideration the news analysis part.We believe that this approach incorporated into existing strategies will encourage quant traders to invest and maximize their profit.The future quant funds will also obviate risks that are seized by unforeseen news events and become more pliable and sturdy.

VI. REFERENCES

[1] Patel, Jigar, Sahil Shah, Priyank Thakkar, and K. Kotecha. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." *Expert Systems with Applications* 42, no. 1 (2015): 259-268.

[2] Patel, Jigar, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. "Predicting stock market index using fusion of machine learning techniques." *Expert Systems with Applications* 42, no. 4 (2015): 2162-2172.

[3] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." *Computers & Operations Research* 32, no. 10 (2005): 2513-2522.

[4] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." *ACM Transactions on Information Systems (TOIS)* 27, no. 2 (2009): 12.

[5] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." *ACM Transactions on Information Systems (TOIS)* 27, no. 2 (2009): 12..

[6] Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." Neurocomputing 55, no. 1-2 (2003): 307-319.

[7] Tsai, C. F., and S. P. Wang. "Stock price forecasting by hybrid machine learning techniques." In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, no. 755, p. 60. 2009..

[8] Yoo, Paul D., Maria H. Kim, and Tony Jan. "Machine learning techniques and use of event information for stock market prediction: A survey and evaluation." In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, vol. 2, pp. 835-841. IEEE, 2005.

[9] Trafalis, Theodore B., and Huseyin Ince. "Support vector machine for regression and applications to financial forecasting." In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 6, pp. 348-353. IEEE, 2000.

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

[11] Bose, Indranil, and Radha K. Mahapatra. "Business data mining—a machine learning perspective." *Information & management* 39, no. 3 (2001): 211-225.

[12] Schumaker, Robert, and Hsinchun Chen. "Textual analysis of stock market prediction using financial news articles." *AMCIS 2006 Proceedings* (2006): 185.

[13] Schumaker, Robert, and Hsinchun Chen. "Textual analysis of stock market prediction using financial news articles." *AMCIS 2006 Proceedings* (2006): 185.

[14] Wang, Yanshan, and In-Chan Choi. "Market index and stock price direction prediction using machine learning techniques: an empirical study on the KOSPI and HSI." *arXiv preprint arXiv:1309.7119* (2013).

[15] Zhai, Yuzheng, Arthur Hsu, and Saman K. Halgamuge. "Combining news and technical indicators in daily stock price trends prediction." In *International symposium on neural networks*, pp. 1087-1096. Springer, Berlin, Heidelberg, 2007.