International Research Journal of Computer Science (IRJCS)
Issue 04, Volume 6 (April 2019)

ISSN: 2393-9842
www.irjcs.com

REAL TIME PREDICTIVE ANALYSIS OF INDIAN STOCK MARKET USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

Jay Kakkad
Computer Engineering,
K.J Somaiya College of Engineering, Mumbai, Maharashtra
jay.kakkad@somaiya.edu

Saurabh Makwana
Computer Engineering,
K.J Somaiya College of Engineering, Mumbai, Maharashtra
saurabhkumar.m@somaiya.edu

Riya Shah
Computer Engineering,
K.J Somaiya College of Engineering, Mumbai, Maharashtra
riya.vs@somaiya.edu

Shweta Chachra
Computer Engineering,
K.J Somaiya College of Engineering, Mumbai, Maharashtra
shweta.chachra@somaiya.edu

Manuscript History
Number: **IRJCS/RS/Vol.06/Issue04/APCS10093**
Received: 11, April 2019
Final Correction: 19, April 2019
Final Accepted: 25, April 2019
Published: April 2019
Citation: Jay, Saurabh, Riya & Shweta (2019). Real Time Predictive Analysis of Indian Stock Market Using Machine Learning and Natural Language Processing. IRJCS:: International Research Journal of Computer Science, Volume VI, 184-188. **doi://10.26562/IRJCS.2019.APCS10088**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract— In a financially volatile market, as in the case of asset market, it is important to have a very precise prediction of the upcoming trends to take advantage of the market changes. This requires highly advanced machine learning algorithms with factorization of human sentiments. The value of an asset is highly gullible due to factors such as market news and human behaviour, which can instantly increase or decrease the asset price. Therefore, the issue becomes that of buying or selling asset at an asset exchange at right moment for generating profit. This aspect has attracted researchers for years and has led to creation of various different algorithms that can predict the outcome in this nonlinear volatile market. This wave of algorithm in Artificial Intelligence, mainly machine learning are still being tried to tackle the above problem. It is also very well known that news articles and related information have great impact on the asset price and its trends. Hence, we aim to combine these two approaches in an attempt to understand the relationship between the attributes in order to yield a predicted output with great accuracy.

Keywords— Asset, Database, LaGrange, Natural Language Processing, pandas, Regression, Stopwords, Stock Market.

I. INTRODUCTION

Stock exchanges are financial institutions, which set up a platform for trading of securities. These securities help own parts of a corporation without sometimes taking part in the decision making or running of the business. Movement in price of each asset traded is the result of human and statistical quotient and is constantly changing. The value of asset, in the beginning is based on the assets of company, but later becomes dependent on the value that the business can produce and the trust people have in the company. The greater the number of people interested in buying shares, the greater the increase in the value of the company asset. Machine Learning has already been used in various algorithms for asset trading and has been proven a great helping hand for financial predictions. Existing technologies for trading mainly High Frequency Trading (HFT) is accessible and used by only handful people around the world, mainly due to one, it's extremely high cost of operation and maintenance and two, it's been banned in a few countries due to its disruptive behaviour. The problem hence arises to create a smart yet economical platform for predicting value of shares that can be utilized by general traders who trade in a relatively low volume transaction.

A few questions arise while tackling this problem such as

- What are the optimal features involved to analyse nonlinearity?
- Is the problem defined as classification or regression?
- How to integrate algorithms into the software architecture to efficiently predict the future?

We propose a novel method of predicting asset market prices based on the integration of sentimental value of the news article related to the asset data along with the intermediate prediction made by the Support Vector Regression (SVR) machine learning model to achieve more accurate results. The machine learning model will train and fit the training asset data in such a manner so that it minimizes as much error as possible. On the other side, the natural language processing model will determine the sentimental polarity of the news data predicts most of the variation caused by such news article and its corresponding impact on the asset price. Hence, we aim to efficiently determine the predicted value of asset considering all the above-mentioned factors and attributes.

II. METHODOLOGY

We intend to combine our system as per the high-level design provided in Figure II.1.

A. Database

Due to the advantages of flexibility and accessibility provided by cloud, we have decided to use it to develop our model's backend.

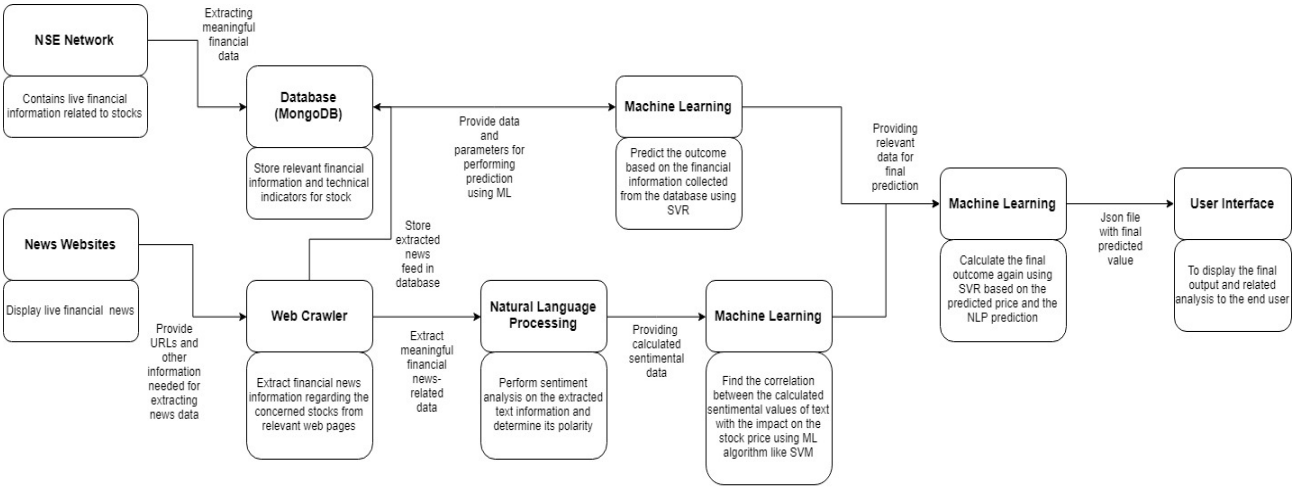


Fig. II.1 Proposed System Model

After having surveyed plethora of different database systems, we have decided to use MongoDB. The information available online and the data gathered from it better suits the unstructured format provided by cost effective MongoDB where we have the flexibility to create and merge numerous clusters. It helps us create logical relationships between records and data tables and allows multi user and concurrent access so that ease of use for the developer is also better. The developer can modify the load sharing capacity as per the terms required. We have used Google cloud platform's cluster due to its highly developed security mechanism and load sharing mechanism. Python, connected via our database server, is used to convert our unstructured data into structured pandas data frame format.

B. Web Crawler

Our prediction-processing model is based on combination of two parameters, news based on textual and asset pertaining to statistical parameters. To obtain information for the above we had to crawl through World Wide Web platform. Our model uses News-Please and NewsPaper free python packages. API NewsPaper gives us the ability to crawl numerous URL links on a website whereas API News-Please gives the ability to scrape out content from given URL links. One of the most useful feature of NewsPaper API is its ability to not scrape the similar news content on its next run. Use of another API HtmlDate has been used to scrape out time of publish on that content as it's crucial to create a timeline for this content and then relate them to our asset. Our Natural Language Processing model uses only recognized certified textual inputs hence use of only well recognized news source have been taken as input parameters. All scraped news gets converted into a json file format which is then stored to our database. Apart from news we'll also be requiring asset prices but to formulate a model on all assets would require a lot of infrastructure hence three NSE assets namely Pfizer, SunPharma and Dr. Reddy are the assets on which we have built our model. We scrape our asset data on inter-day basis. This is possible through pandas_datareader API where data is being fetched from yahoo finance in pandas dataframe format and then converted to json for storing it in to our MongoDB database server.

C. Selection of Technical Indicators

This application is about predicting asset market prices, hence a greater research had to be done on how financial markets operate and what are the indicators being used to determine a result on a human level. Our selection on use of Moving Average, Exponential Moving Average, Relative Strength Index, Volume Weighted Average Price and stochastic oscillator (STOCH) was based on the fact that assets which are under observation are liquid assets with large trading volumes. These indicators will not perform in expected manner if same analogy is implemented on low volume trading asset. Selection, pre-processing and cleansing of data is extremely important for maximizing algorithm's potential. Moving average and exponential moving average determines the trend of the market on a broader perspective which is helpful in excluding daily market oscillations. RSI helps us determine momentum oscillation at which speed and price changes which helps us determine whether particular asset is overbought or oversold hence comparing bullish vs bearish price momentum against assets price. RSI helps us in analysing short term out comings of an asset whereas moving average analyses assets long term out comings. Our SVR model is then used to determine weightage of each indicator

D. Support Vector Regression

Support Vector Regression is used as our core algorithm. It gives the ability of classification between UPTREND and DOWNTREND along with pin point measurement accuracy of a regression model. This model was selected based on the thought process of predicting inter-day closing value of an asset. SVR has incorporated its kernel model from its predecessor Support vector machine model, which is structured only for classification as shown Fig II.D.1.

$$f(x,\omega) = \sum_{j=1}^{m} \omega_j g_j(x) + b$$

Fig. II.D.1 Elementary Kernel equation for Support vector regression

Lagrangian model is used for convex optimization which helps it evaluating and improving training model as and when it grows. This model is shown through Fig II.D.2.

$$L_\varepsilon(y, f(x,\omega)) = \begin{cases} 0 & \text{if } |y - f(x,\omega)| \le \varepsilon \\ |y - f(x,\omega)| - \varepsilon & \text{otherwise} \end{cases}$$

Fig. II.D.2 LaGrangian Optimization

Each supervised learning model has its constraint model and as support vector regression is hybrid model of classification and regression Fig D.3 is the constraint model defined which is then used in developing final lagrangian optimization model

$$\begin{cases} y_i - f(x_i,\omega) \le \varepsilon + \xi_i^* \\ f(x_i,\omega) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1,\dots,n \end{cases}$$

Fig. II.D.3 Loss Function

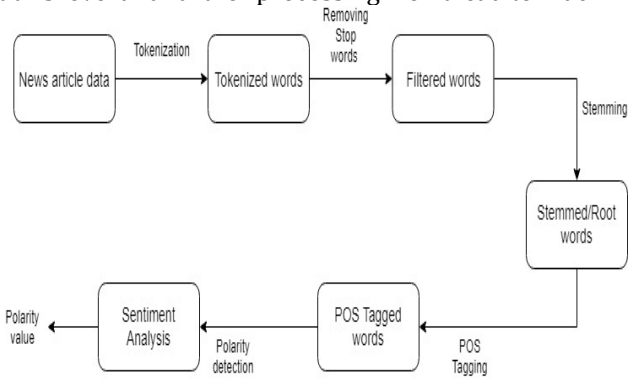
After solving LaGrange's convex optimization problem alpha and alpha* is respectively found through which in case of a coded format if found through numerous iterations after which it is deployment ready.

$$f(x) = \sum_{i=1}^{n_{sv}} (\alpha_i - \alpha_i^*) K(x_i, x)$$

Fig. II.D.4 Final Support vector equation format after Lagrangian Optimization

E. Natural Language Processing

News data collected from the relevant financial sites have meaningful information stored in them in the form of the news article's title, description, content, date and so on. Using date as a key for merging, all articles for the same day are merged together and used for further processing. Then, the standard text processing methods are applied on the selected text. We first tokenize the text, especially the long text in the content of the news article into token of words. By this, we aim to focus on the words which have the potential to yield insights. Rest of the filler stop words like "a", "an", "the" are filtered out in the next step. These words are then passed through the process of stemming where root or stem word is considered for each and every word. Each of these stemmed words is then tagged using the "Part of Speech" tagging process. Finally, we apply two types of polarity detection algorithms for sentiment analysis to calculate the sentimental output value of the processed words. We also apply these algorithms directly on the title and description text at the sentence level since much of the sentimental information is still retained at this level and further processing won't lead to much more accurate results.



NLP Module

Fig. II.E.4.

TextBlob and VADER Sentiment analysis are the two techniques we aim to apply in order to calculate the polarity of the given text. TextBlob would give us a value ranging from -1 to +1, where -1 represents negative polarity, that is, information which is not well-received or as in our case some indication that might lead to the decrease in the asset prices whereas +1 indicates vice-versa. It also helps in calculating the subjectivity of the text which helps us to decide the validity of the information processed. VADER Sentiment Analysis on the other hand would indicate polarity ranging from 0 to 1 on four different levels, that are positive, negative, compound and neutral. Here, the magnitude of the value indicates the intensity of that sentiment.

F. Machine Learning

Each asset price would be indicated with a label to predict the direction of the asset price, that is, whether up or down for that particular day. Based on the previous historical asset data we collected, we will label the classes in binary form to indicate the growth or decline of the asset price for that particular day. This along with the polarity values we calculated using the Natural Language Processing module will serve as the training dataset for our Machine Learning Algorithm. We found out that the Support Vector Machine(SVM) algorithm would work the best in our case to predict the labels given the test data information. Finally, again all the calculated sentimental polarity values along with the labels are passed to the SVR algorithm, which will automatically adjust its coefficients in order to fit the data for regression. The final output will be a prediction value based on the given input of opening day asset price which will be stored in a json file. Database the above section says how to prepare a subsection. Just copy and paste the subsection, whenever you need it. The numbers will be automatically changes when you add new subsection. Once you paste it, change the subsection heading as per your requirements.

G. User Interface

This module is used for interacting with the end user and displaying meaningful information calculated by our algorithms to aid financially illiterate people with the help of data visualization techniques. We have prepared a website using latest web development technologies and implemented prediction charts so that it is easy to see the results. We have also included the latest trending news so that the users can see the type of news that reflects the price change in the market. The user interface has been developed such that it has improved efficiency and performance with ease of usability, familiarity and consistency for the benefit of the user. We hope that the end user is able to take complete advantage of our application and reap the benefits of investing in the market.

Database. The above section says how to prepare a subsection. Just copy and paste the subsection, whenever you need it. The numbers will be automatically changes when you add new subsection. Once you paste it, change the subsection heading as per your requirement.

III. CONCLUSIONS AND FUTURE SCOPE

Stock market prediction depends on a number of factors and has a non-linear relation which we tried to predict using our methodology. The sentiment analysis of the news article helps to predict the trends in the market along with their impact whereas the SVR Machine Learning algorithm takes historical data of the stock into consideration along with other technical indicators which vary from stock to stock. Although both the natural language processing and machine learning part require data preprocessing and some changes needed to be done to the algorithm for each of the stocks for optimization purposes. For future work, one can increase the scope of this project by taking more number and variety of stocks into consideration. Also, one can do a comparative analysis for different approaches used for Machine Learning as well as Sentimental analysis of news feed data to better understand the impact of each algorithm on the prediction of the stock price and determine which combination works the best for a particular stock. Furthermore, one can also incorporate several changes in the SVR as well as NLP algorithms and do a comparative study in them to better understand the nature of each algorithm on stock market prices

ACKNOWLEDGMENT

The [1],[3],[4],[5],[7],[8] have given comprehensive understanding on working of various machine learning algorithms on stock market. In depth knowledge of mathematical functioning of support vector regression model was obtained from [2]. Knowledge and understanding for creation of NLP model was inferred from [4]. A superficial understanding of working of financial market was obtained from [6] and [10].

REFERENCES

1. Shom Prasad Das, Sudarshan Padhy, (March 2012), Support Vector Machines for Prediction of Futures Prices in Indian Stock Market, International Journal of Computer Applications,(0975-8887)
2. Alex J. Smola , Bernhard Schölkopf (November 2004), A tutorial on support vector regression, Statistics and Computing 14: 199–222, 2004
3. Qasem-al-Radaideh, Adel Abu Assaf, Eman Alnagi, (December 2013), Predicting Stock Prices using Data Mining Techniques, International Arab Conference on Information Technology, pp.1-8
4. Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao,(June 2016),Stock Trend Prediction using News Sentiment Analysis,International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 3(0975–9646)
5. Shri Bharthi . Sv,Angelina Geetha,(June 2017),Sentiment Analysis for Effective Stock Market Prediction,International Journal of Intelligent Engineering and Systems(2185-3118)
6. H. White, Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, ser. Discussion paper- Department of Economics University of California San Diego. Department of Economics,University of California, 1988.
7. E. W. Saad, D. V. Prokhorov, and D. C. Wunsch, "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks," IEEE Transactions on neural networks, vol. 9, no. 6,,pp. 1456–1470, 1998.
8. R. Araújo, A.L. Oliveira, S. MeiraA hybrid model for high-frequency stock market forecasting Expert Syst Appl, 42 (8) (2015), pp. 4081-4096
9. S. Choudhury, S. Ghosh, A. Bhattacharya, K.J. Fernandes, M.K. TiwariA real time clustering and SVM based price-volatility prediction for optimal trading strategy Neurocomputing, 131 (1) (2014), pp. 419-426
10. M.-W. Hsu, S. Lessmann, M.-C. Sung, T. Ma, J.E. Johnson Bridging the divide in financial market forecasting: machine learners vs. financial economists Expert Syst Appl, 61 (1) (2016), pp. 215-234
11. NewsPaper API (https://newsapi.org/)
12. https://www.moneycontrol.com/
13. https://www.bloombergquint.com/
14. https://www.livemint.com/
15. https://economictimes.indiatimes.com/
16. https://in.finance.yahoo.com/
17. https://www.mongodb.com/

