International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Predition Model for Stock Price on Big Data Analytics

Thin Thin Swe, Phyu Phyu, Sandar Pa Pa Thein

Lecturer, Faculty of Information Science, Universtiy of Computer Studies, Pathein, Myanmar

How to cite this paper: Thin Thin Swe | Phyu Phyu | Sandar Pa Pa Thein "Predition Model for Stock Price on Big Data Analytics" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1444-1446, **https://doi.org/10.31142/ijtsrd26660**

IJTSRD26660

ABSTRACT

Prediction in the stock market is very challenging in these days. Large datasets available from Twitter micro blogging platform and widely available stock market records. Machine learning was employ to conduct sentiment analysis of data and to estimate for future stock prices. The relation between sentiments and the stock value is to be determined. A comparative study of these algorithms: Multiple linear Regression, Support Vector Machine and Artificial Neural Network are done.

KEYWORDS: *Stock Market; Sentiment Analysis; Multiple Linear Regression (MLR); Support Vector Machine (SVM); Artificial Neural Network (ANN)*

I. INTRODUCTION

The Stock Market Is Always One Of The Most Popular Investments Due To Its High Profit. Recent Years Have Shown Not Only An Explosion Of Data But Also Widespread Attempts To Analyse It For Practise Reasons. Scientists And Computer Engineers Have Crated Special Term "Big Data" To Name This Trend. Main Features Of Big Data Are Volume, Variety And Velocity. Volume Refers To The Amount Of Data, Variety Refers To The Number Of Types Of Data And Velocity Refers To The Speed Of Data Processing.

There Are Many Reasons To Grow Of Big Data. One Of The Main Reason Is Computer Systems Start To Be Used In Many Sectors Of The Economy From Governments And Local Authories To Help To Financial Sector.

The Analyses Of Information That Is A By-Product Of Different Business Activities By Companies Can Lead To Better Understanding The Needs Of Their Customers And Preditions Future Trends. It Was Previously Reported In Several Research Papers That Precise Financial Markets [1].

A. Big Data

There are several definitions what Big data is , one of them is following: "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse." [2] This definition emphasizes key aspects of big data that are volume, velocity and variety [3]. According to IBM reports [4] everyday "2.5 quintillion bytes of data" is created. These figures are increasing each year. This is due to previously described ubiquitous access to the Internet and growing number of devices. Data is created and delivered from various systems operating in real-time. For example social media platforms aggregate constantly information about user activities and interactions e.g. one of most popular social sites Facebook has over 618 million daily active users [5]. Output rate of the system can be also important when nearly real-time analyses are needed.

But big data is not only challenging but primarily creates opportunities. They are, among the others: creating transparency, optimization and improving performance, generation of additional profits and nothing else than discovering new ideas, services and products.

B. Social Media

Twitter is an online news and social networking site where people can communicate in short messages. Twitter is a blend of social media, blogging and texting. Twitter employs a purposeful message size restriction to keep things scan friendly: every microblog tweet entry is limited to 280 characters or less. This size cap promotes the focused and clever use of language, which makes tweets easy to scan and challenging to write. This size restriction made Twitter a popular social tool [4]. Therefore Twitter was chosen for experimental data source for this work on predicting stock market.

II. Predicting Future stock prices

The stock market prediction is difficult since the stock price is dynamic in nature. To reduce the false forecasts of the stock market and increase the ability to predict the market movements. To avoid the risk and the complex in predicting stock price.

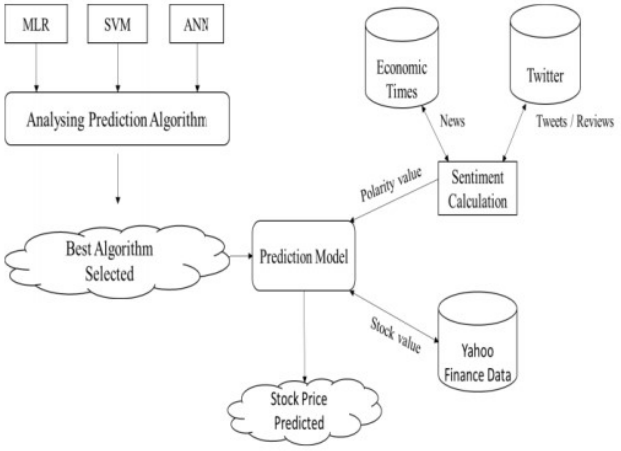

Figure 1: Design of the System

A. Data Collection and Processing

The Yahoo finance data contain BSE data and NSE data. The financial index of BSE is SENSEX and for NSE is NIFTY. This finance data contain many attributes of the stock market. Some of the attributes are opening value, closing value, adjacent close value, min value, max value, volume, dividend, date and time of the day. The key attributes are selected for stock market prediction in this work. The collected data are processed by normalization and used as the input data. For monthly prediction and daily prediction historical data from Yahoo finance is used.

B. Multiple Linear Regression

Regression is a data mining task of predicting the stock price by implementing a model based on one or more attributes. This technique is a compact method of mathematical representation. It represents the relationship between the response parameter and the input parameter in a given prediction model. Prediction of the outcome y from the input parameter x of the form

$$Y = \beta 0 + \beta 1\ x1 + \beta 2\ x2 \ldots + \beta k\ xk \quad (1)$$

y – Outcome, β0 – Intercept, β1, β2, βk – Partial Regression Co-efficient, x1, x2, xk – Input Parameters.

The least square value is determined by the current market price y from the PE ratio and mean of the input parameters

$$y = a + bx \quad (2)$$

C. Support Vector Machine

In machine learning, Support vector machine is supervised learning models with associated learning algorithms to identify the data. This process includes the classification and regression methods to analyse the prediction. It implements the margins adoption to predict the market price. The support vector means the closest hyper-plane and it is equal in distance are identified. The value of support vector y is calculated using the two attribute case .

$$y = w0 + w1\ x1 + w2\ x2 \quad (3)$$

y – Outcome, w0, w1, w2 – three weights to be learned by SVM model, x1, x2 – input parameters

D. Artificial Neural Network

The artificial neural network is deep learning at hidden layer neuron. Artificial Intelligence classifiers are used to train the prediction model to identify the trends in a complicated environment. The input value is assigned with different weights in the input layer. The output of the input layer is fed into the hidden layer neuron as the input. The summation function is evaluated in the hidden layer. It learns every prediction at each node. If any hidden layer neuron is not completed, then it is fed into the summation function.

$$h = \sum wi.xi \quad (4)$$

h – Summation function, ∑wi – Weight of the input, xi – Input value.

The output of the hidden layer is given as input to the output layer where the sigmoid function O is determined using the summation function h.

$$O = 1 / 1+e\ (-h) \quad (5)$$

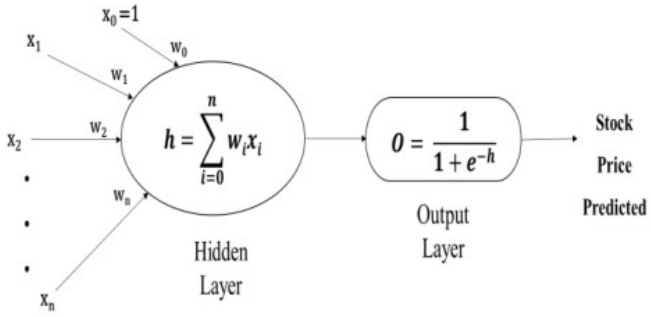

Figure2: Process of ANN

E. Sentiment Analysis

The expression of each individual varies and it is analysed in social media data. The polarity of each word is evaluated. The polarity can be neutral, positive and negative. The correlation between the news and actual price is determined. The Sentiment value in a row S is initial zero. For each word in a row do the Sentiment Calculation.

Sent_Word (word,dictionary) (6)
Sent_Word : compare the word with dictionary.

S = Sent_Word (7)
S – Calculates the value of sentiment anlysis of the data.

III. discussion of result

As it can be observed from presented results, predictions of stock prices depend strongly on choice of training dataset, their preparation methods and number of appearing messages per time interval. Predictions conducted with models trained with datasets with messages containing company stock symbol performs better. It can be explained by the fact that these messages refer to stock market. Tweets with company name may just transfer information, which does not affect financial results.

Another important factor is a choice of preparation of training set. Two methods were used. One of the methods was a manual labeling sentiment value of messages. This method allows to more accurately label training data but is not effective for creating large training sets. The other method was applying SentiWordNet, which is a lexical resource for sentiment opinion mining.

It enabled to create bigger training datasets, which resulted in building more accurate models. Last factor that is important for prediction is number of appearing messages per time interval. Although model trained with datasets with company name were not accurate in comparison to the other datasets, there is bigger number of tweets per time interval. It allowed performing prediction for shorter time intervals, which were not possible for dataset with messages containing company stock symbol. Described methods can be also used with other stock predictions procedures in order to maintain higher accuracy. It is also important to note that stock prediction methods are not able to predict sudden events called 'black swans' [7] .

IV. conclusions

The purpose of this study is to compare the performance of the three prediction algorithms Multiple Linear Regression, Support Vector Machine, Artificial Neural Network in the stock market. The Multiple Linear Regression

algorithm is less developed state which measures the relationship between the stock price and volume. The Support Vector Machine algorithm is a two-class classifier for the learning model.

The Artificial Neural Network is the classification algorithm for deep learning. The result exhibits that the deep learning algorithm performs better than the MLR and SVM. In deep learning algorithm the hidden layer neuron learns in every prediction. Hence the output layer neuron produces the best outcome. Artificial Neural Network is the best predicting algorithm.

References

[1] Z. Da, J. Engelberg, P. Gao: *In Search of Attention*, The Journal of Finance Volume 66, Issue 5, pages 1461–1499, October 2011, doi: 10.1111/j.1540-6261.2011.01679.x

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition, and productivity, McKinsey, May 2011..

[3] Edd Dumbill. What is big data? : an introduction to the big data landscape. http:// radar.oreilly.com/2012/01/what-is big-data.html, 2012 R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] What is Twitter and How does it work? http://www.lifewire.com/ what-is –twitter.

[5] Hagenau, Michael, Michael Liebmann, Markus Hedwig, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context-specific features." In System Science (HICSS), 2012 45th Hawaii International Conference on, pp. 1040-1049. IEEE, 2012.

[6] Khan, Zabir Haider, Tasnim Sharmin Alin, and Md Akter Hussain. "Price prediction of share market using artificial neural network (ANN)." International Journal of Computer Applications 22, no. 2 (2011): 42-47

[7] N. N. Taleb, Common Errors in the Interpretation of the Ideas of Then Black Swan and Associated Papers.