A Survey on Various methods for Stock Prediction using Big Data Analytics

Ashish Awate
Department ofComputer Engineering
SVKM's Intitute of Technology
Dhule,India
ashish.awate87@gmail.com

Bhushan Nandwalkar
Department ofComputer Engineering
SVKM's Intitute of Technology
Dhule,India
nandwalkar.bhushan@gmail.com

***Abstract*—this paper is a survey about various methods in Big Data Analytics implemented to predict future stock trends or stock price. Initially paper starts with exploring the concepts of both the Big Data and Big Data Analytics. The paper surveyed total six different papers which consist of different technique to predict stock price. We survey this paper on the basis Theme, Proposed Method, Experimentation, Results/Advantage, and Limitation. These papers illustrate different methods to predict future stock price. The survey concludes with predictive Big Data Analytics is more suitable technique for stock prediction and also dataset must be large enough for train and test.**

Keywords—Big Data, Big Data Analytics, Stock Prediction, Data Mining, Machine Learning.

I. INTRODUCTION

This paper has two parts first comprises of Big Data and Big Data analytics and second part has Stock and stock prediction. Let's see the first the Big Data and big Data analytics.

Big Data term generally used for data sets which are huge and having large volume. It is applicable for both structured and unstructured data. This data is too complex to be handled by traditional data processing software or database, so it required some new software and techniques to performed desired task in bounded time. But it is important to state that the term 'Big Data'[1] is always paly role when data is huge i.e. in gigabytes or terabytes rather that a small data set can also termed as Big Data depend upon the context for which we used it.

The main difference between the traditional Relational database and Big Data is that, Big Data uses multiple processor for multiset information rather that stick to single node processing for multiset information.

A. Need of Big Data

We are living in the world where day by day the digitalization and internet is touching each aspects of our day today life and as we know each use of internet and digital device producing some or more amount of data depends on its use. These are raw data which are supposed to be processed and bring out desired pattern predictions or conclusion as per need. But the traditional ways such as relational databases, spreadsheets etc. of processing are unable to handle such huge amount of data and bring out some conclusion in stipulated time. As a remedy of this problem Big Data came out with solution.

B. Big Data Charasterastics

Big data has four characteristics also well known as 4 V's of big Data namely Volume, Velocity, Variety and Value. Other than this Veracity, Validity and Volatility also consider as characteristics of big Data [1].

1) Volume : The amount of data used for processing is termed as Volume. Depends on the requirement of application the data can be huge or small.

2) Velocity : The speed in chich the data procesed is termed as Velocity. To manipulate the time of processesing this characterastics is important.

3) Variety : Big Data can deal with both structured and Unstructured data. So we need to discriminate both types of data. So the different kinds of data came out of processing is termed as Variety.

4) Value : Categorization of usage of huge data and small data which then combined to complete particular task is termed as Value. To offer the quality analytics Value uses the Volume and Variety of data to be operated.

C. Significance of Big Data

With help of numbers of processors,to process the large amount of data in pallely and efficiently is the main aspect of Big Data. In short span of time Big data provides the output by manipulating, computing, analyzing any amount of information.

D. Applications of Big Data

Education, government, Insurance, Media, Healthcare, manufacturing, international development, science, research, etc. are the fields where Big Data have significance impact. Each field depend on the volume and variety of data it produce has different setup and method to implement Big Data.

E. Challenges in Big Data [5]

As there are various significance of Big Data, yet to implement and deploy Big Data in various field we might face good numbers of the challenges. They are enlisted as

- Rapid change and continuous shift in technology
- Scarcity of expert human resource to process the data
- Threats on security and privacy of data
- As the data preserved on cloud it introduces complexity in implementation.
- Difference in understanding and actual implementation

II. Big Data Analytics

Big Data analytics is the term proposes in harmony of Big Data to find the "hidden patterns, unseen correlation, business decisions, user preference, drifts in market, sentiment on social network, and unknown statistical associations".

A. Significance of Big Data Analytics

To get the specific yield from the huge amount of data within the stipulated time period is the main role of Big Data The significance of Big Data analytics are as follows.

1) Reduction in implementation Cost : Due to use of tools like Hadoop for mapping the data and which considerably reduce the volume of data on cloud storage impact reduction of implementing cost.

2) Fast and Improved Decision Making: Big Data Tools analyzed information quickly for making faster decision and analytical toolwith the improved esult can help to predict the future treds and one can take accurate decision.

B. Types of Data Analytics

Descriptive, Diagnostics, Predictive and Prescriptive analytics are the four types in which Big Data Analytics are classified.

1) Descriptive Analytics : The question of 'what happened' the answer resides in Descriptive analytics. For example, a education provider will learn how many students were admitted in last year; a shopkeeper – the average weekly sales volume; a manufacturer – a rate of the products returned for a past month, etc. Descriptive analytics manipulates raw data from multiple sources to provides valuable perceptions into the past. These discoveries simply signal that what to do, without explaining why. Due to this companies vastly rely on data do not content themselves solely on descriptive analytics only, rather they prefered to combination of more than one type of data analytics.

2) Diagnostics Analytics : This method of analytics provides the answer of question "why something happened*"* considering historical data into account. Diagnostic analytics finds patterns and dependencies inside the data by drilling it down. As it provides the insight of particular problem, so companies attract towards diagnostic analytics, But to implement Diagnostic analytics company must have detailed information, otherwise it will gives create wrong pattenr and dependencies and it become time-consuming.

3) Predictive Analytics : This method of analytics provides the answer of question " what is likely to happen?". It predicts the future trends by taking findings of descriptive and diagnostic analytics into consideration to spot current trends, groups and exceptions. So it is a valuable tool for forecasting. But it is important to note that predictive analytics provide just an estimated prediction dpens on quality, amount and stablity of data. So one must take care while treating the data and also contineous optimization is also needed.

4) Prescriptive analytics : This method of analytics provides the answer of question "prescribe what action to take? " So it will help the companies of firm to eradicate a future problem or take full advantage of a promising trend. An example of prescriptive analytics from our paper portfolio: a multinational company was able to identify opportunities for repeat purchases based on customer analytics and sales history. It also required external information with historical data as per the nature of statistical algorithms used. Also, prescriptive analytics uses machine learning, business rules and algorithms. So, it is important company must do homework about required efforts vs. an expected added value before adopting prescriptive analytics[3].

C. Application Areas of Data Analytics:

Data analytics makes a firm, institute or company intelligent to take decision and decide what to do next? There are many application and use of Data analytics. Some of them are in sectors like Government, Medicines, Education, Military, Manufacturing, and Logistics. But for this survey paper focus is on application of data analytics in financial sector. They are as follows

1) Fraud Detection: Fraud is term came into picture when some person or machine unauthorized or unauthenticated way access the system and stole the data or person provides wrong information as identity and get financial access of firm. In both situations the concern firm or company may lose money. So it is mandatory to detect fraud before it happens, to do this data analytics provides good tool and indication method to detect fraudulent access or fraud mechanism from gaining the access.

2) Economy-Level Prediction: Every country depends on the economical structure of it and on the basis of that government plans the different projects for the people. But planning needs the economic level prediction for next financial year. To do this data analytics methods plays vital role predict economic level of country based on current happening in country. Same concept can be applied to company.

3) Stock Prediction: Stocks trends and price are most dynamics in nature means it changes every tick of clock since market opens to market get closed. Most of people do

trading to earn money and but if someone wrongly predict the price of stock it may cause great deal of money. To avoid that one must predict future price of stock correctly to do this predictive analytics is the best way to do this.

III. Stock Market Prediction

In share marketing for investment we have to use proper analysis of the stock. There are two different methods for stock prediction first is fundamental analysis and second is technical analysis. In fundamental analysis of stock we have to consider different parameters of stock like company capital, profit loss, future business of organization, current business growth in domain, and government policies towards organization domain, contribution of stock in share market, reserve funds and management power towards business decision. In Other hand technical analysis is totally different than fundamental analysis. In technical analysis we consider ups and downs in prices of stock in share market using different charts. In technical analysis to predict sentiment and future potential patterns we have to consider historical prices of stock and volume of stock. Most of the people uses different methods for doing technical analysis like chart patterns or indictors and oscillators. A specific strategy is used for to decide Price of stock in technical analysis these strategies may be to find out specific trends in market and if trend is stable then use these trends to predict future stock analysis and prices goes up and down as per these trends patterns.

A. Market Stock Types

Market is location where stocks of companies are traded by investors. Initially share market has two parts, in first part new stocks are introduced by companies called as Initial Public Offerings. In second part, stock investors do the trading over stock. Stocks can be classified into different categories on various parameters like size of the company, dividend payment, industry, risk, volatility, as well as fundamentals**.**

B. Big Data Vs Stock Prediction

Extremely large data or big data set have three properties volume, velocity, and verity of data. Financial organization and traders can find out different patterns for stock prediction and good decision making through extracting information from big data using big data analytics.

C. Survey on Stock Prediction Method

1) *Stock Transaction Analysis System based on Hadoop and Capital flow[2]*
   a. *Theme:* To test if the stock transaction behavior can be detected by using data processing technique on a synthetic index of capital flow.
   b. *Proposed Method:* Author has proposed logical architecture and framework to analysis of system also they introduced the prototype to showcase how they implement their idea.
   c. *Experimentation:* They adopt hierarchical structure and divide their implementation in two parts. In first Part is Hardware part they use of 4 nodes of Hadoop cluster which are implemented in the VM. And second is software part in that they use Web Crawler to get transaction data from all kinds of public information portals. The collected data fed into JAVA server, which is used for booting and maintaining Hadoop cluster to complete the data processing and calculation of stock analysis.
   d. *Result / Advantage :* They conclude by stating that the prototype system is able to provide better trade-off among accuracy, availability and flexibility
   e. *Limitations:* This technique is purely on statistical approach so it can test based on daily transaction data. The used method for data analysis method and scale of data both are insufficient to predict practical stock transaction.

2) *Stock Price Prediction Using Data Analytics[3]*
   a. *Theme:* Opening Stock price prediction using data analytics
   b. *Proposed Method:* Famous Data analytics methods R and Python were used along with artificial intelligence method to train and test tweets data
   c. *Experimentation:* As training data they have used previous 9 year data of 50 stocks of Nifty along with recent Tweets for same stocks to do sentiment analysis. For R platform they used multiple techniques like Arima, Holt winters, neural networks (Feed forward and Multi-layer perceptron), linear regression and time series are implemented to forecast the opening index price performance in R. The methods such as Multi-layer perceptron and support vector regression are implemented in Python. They measured accuracy by comparing actual price of stocks by using 2-3 years forecast of result in R and for python they do only 2 months of forecast results.
   d. *Result / Advantage:* By using actual intact raw data they got 1.81598342% of mean absolute percentage error for feed forward neural network which is least amongst used methods and for the linear model with polynomial trend the mean absolute percentage error is 11.32847594% which is highest amongst all.
   e. *Limitations:* Results are only for opening price of stock the variation and closing price not predict. And also prediction was average for a whole month.

3) *Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction[4]*

a. *Theme:* Analysis of data on social media services to predict future stock prices.
b. *Proposed Method :* To predict the stock price authors has used the machine learning algorithms for sentiment classification of large data set tweeter of Apple Inc company in distributed environment of Map-Reduced environment
c. *Experimentation:* To implement system author has used two types of datasets, first dataset contains tweets containing name of the company and the second one contains stock symbol. The predictions were made for Apple Inc. The three months Data was collected and processed for analysis. Two methods were used first one is manual Labeling of sentiment value of message and second one is lexical resource for sentiment opinion mining called SentiWordNet.
d. *Results/ Advantage:* In comparison with both dataset, dataset containing stock symbol perform better than datasets with company name. Same Method SentiWordNet gives better result than manual labeling.
e. *Limitations:* As stock symbol gives better result but it is not big enough to produce accurate result and company dataset is big enough but it take long time to predict result accurately.

4) *Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction:[5]*
a. *Theme :* Predict the future stock price based on live tweeter data using real-time sentiment analysis
b. *Proposed Method:* Author has used historical data to classifying model. They provided perfect data as input to model since the prediction are based on real-time basis. Next they process this huge data using Spark streaming, Twitter API was used to fetch data and for analysis part Apache Fumes were used.
c. *Experimentation:* Implementation was done by, Lambda Architecture: used for stream analysis with three layer namely batch layer, speed layer, and serving layer. The Twitter API fetch the data and puts in HDFS(Hadoop Distributed File System). The data collected were Time Series data we need to break down it, to do these authors has used information visualization (InfoVis) and visual analytics (VA). Standford Core NLP's RNN component was used to do Twitter Sentiments analysis. MLlib component of spark with Naive Bayes classifier were used for analysis of History Data. Stanfords CoreNLP API used for Real Time Sentiment Analysis. Training dataset consist of 56000 tweet of last 13 years taken from Yahoo finance site for Actual Stock price of Google, Microsoft and Apple. For testing purpose 200 days twitter data taken into consideration.
d. *Result / Advantage:* This paper provides good indication that sentiment analysis done by extracting public mood from Twitter feeds can be used to forecast movements of individual stock prices.
e. *Limitations:* Method work on real-time data and it is not necessary the desired data is available all the time. Also it uses Spark's MLlib which is not fully matured method in Machine Learning.

5) *Using Social Media Mining Technology to Assist in Price Prediction of Stock Market[6]*
a. *Theme:* To predict the small cap stock price using social media mining technology.
b. *Proposed Method:* In this paper authors has proposed the three step method. In first step by using Topical Crawler they gather the information from social media followed by labeling the data as a preprocessing technique and fetch this data sentiment analysis algorithm to evaluate based on segment associated with each stock. Finally by using SVM model the prediction of stock price has been carried out.
c. *Experimentation:* They used Yahoo and Google API for gathering data and kept it in HDFS. The Topical Crawler used for as social platform to collect stock comment data. For segmentation they used tool called ICTCLAS and for labeling Net+ sentiment dictionary were used. Finally they calculate Sentiment Index (SI) and Sentiment Discrepancy Index (SDI). They used two models for stock prediction namely SVM_Sentiment and SVM. Total 600719 stock sentiment dataset were used.
d. *Result / Advantage :* They stated that SVM model containing segments index of each stock predict much closer to actual price of stock

e. Limitation: This paper only considers small cap stock price prediction. Also as it based on segmentation so if preprocessing of data must be cautiously done.

6) *NSE Stock Market Prediction Using Deep-Learning Models[7]*
   a. *Theme:* Prediction of NSE and NYSE market stock price using historical data using Deep learning methods
   b. *Proposed Method:* Author has used Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Convolution Neural Network (CNN) of Deep Learning for stock market price prediction based on previous years data available. NSE (National Stock Exchange) of India and NYSE (New York Stock Exchange) closing price has been taken into consideration. Training of network has been done for one stock prize and tried to predict five different stock prices from both NSE and NYSE. The network was trained with the stock price of a single company from NSE and predicted for five different companies from both NSE and NYSE.
   c. *Experimentation:* For experimentation authors have chosen 3 different sectors of market namely Automobile, Bank and IT sector. They consider closing price of each stock for training and prediction. The first Dataset taken for training was TATAMOTORS from automobile sector. The dataset was build over 4861 days and then extracted data normalized using MIN-MAX normalization method. They set different window sizes ranges from 50 to 250 to predict future days price of stock that 10, 20, 30 and 40. According to combination of window size and predicted days they calculated MAPE (Mean Absolute Percentage Error). After this window size is fixed to 200 and 10 days prediction is set as it gives good result and test the outcome for stocks HCL Technologies, Maruti and Axis Bank of NSE and BANK OF AMERICA (BAC) and CHESAPEAK ENERGY (CHK) from NYSE test dataset which was taken into consideration Finally the comparison of result ARIMA model was selected
   d. *Result/ Advantage:* It has been observed that CNN is outperforming the other models. The window size of 200 resulted minimum error than other window sizes. Minimum MAPE is obtained with window size 200 for 10 days prediction.
   e. *Limitation:* This work hasn't explored the advantage of using a hybrid network which combines two networks to make a model for prediction.

IV. Discussion

In this paper, the concepts of Big Data and Big Data Analytics were studied. After this some good methods for Stock Prediction using various approaches is taken into consideration. First paper started with definition, need importance, need, Challenges applications of Big Data were discussed. Followed by Big Data analytics with definition, importance, its types and its application areas were discussed. In the section of Big Data analytics types all four techniques namely Descriptive, Diagnostic, Predictive and Perspective and its probable utilization have been included. Applicability of Big Data Analytics was wide spread but this paper mainly focuses applications like Economy-Level Prediction, Fraud Detection and Stock Prediction are included. Next section comprises of the concept called Stock Prediction in that what is stock, its nature and broad types of stock is included. Followed by how Big Data analytics suited best in predicting future stock price is discussed. To do this survey of total 6 papers having different approach towards predicting stock price were conducted which has flavour of Big Data Analytics.

Papers surveyed in prediction of stock section, there are some findings

- Survey gives the indication that out of four methods of data analytics predictive data analytics is best suited method for future stock prediction or for understanding the market trends.
- To get more accurate results the training data must be large enough and must contain relevant attributes
- To get proper dataset as training input it is important to apply extraction of data carefully and apply normalization if needed
- The single level labeling can be done on sentiment dataset for categorization, to achieve more accuracy multi label learning can be inculcate for more accuracy.
- Mood of people called sentiments and historical data both can be trained to achieve accuracy in prediction.
- It is important to keep data intact and the used dataset must be authenticated.
- There are many resources available for historical data and Sentiments data can be collected from any micro blogging site but survey suggested Yahoo Finance is good for Historical data and Twitter is good for sentiment collection. The tweet data can be gathered either from Search Twitter API or Stream Twitter API.
- For storing this huge dataset HDFS is good
- Survey also suggest to that applying more than one predictive analytics method like regression, clustering, classification together gives least amount of mean absolute error and produces good

Performance factors for stock prediction model. So one can used hybrid approach to implement accurate result.

- To increased effectiveness of machine learning and deep learning methods for predicting stock price one can used optimization algorithms inspired by genetics can be used [12].
- It is also important to note that not only train data must be large but also testing data must be large enough to check accuracy of model.
- For comparison between train and test data ARIMA model can be used.

V. CONCLUSION

This paper starts with stating concept of Big Data, Big Data Analytics. Then the various types to do Data Analytics and classical and technical Stock Price Prediction techniques then included in this paper. Then survey of different methods that can be applied along with Big Data Analytics to predict Stock Price prediction is included. A flow and content of this paper is described in last part of paper which is Discussion. This survey illustrates possibly all techniques related to Big Data Analytics for Stock Prediction. From this survey it is clear that predictive data analytics techniques is best for predicting future trends in market or stock price. Survey concludes large enough dataset containing combination of historical and sentiment data to create test data then extracted and normalized with proper approach and fetch the data to appropriate method of predictive analysis techniques, machine learning method or deep learning method and finally compared it with good number of test data and compare result with appropriate method will enhance the accuracy of stock prediction.

REFERENCES

University" – Computer and Information Sciences 30 (2018) 431–448Big

[2] Feng yu, Sheng-lin Cao, Shu-cheng Huang, Peng-fei Gan, "Stock Transaction Analysis System based on Hadoop and Capital flow". 2018 Sixth International Conference on Advanced Cloud and Big Data 2018. pp. 54-59.

[3] Shashank Tiwari, Akshay Bharadwaj, Dr. Sudha Gupta, **"**Stock Price Prediction Using Data Analytics". 2017 International Conference on Advances in Computing, Communication and Control Dec 2017. pp 1-5.

[4] Michal Skuza, Andrzej Romanowski Sentiment "Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction". Federated Conference on Computer Science and Information Systems 2015 pp. 1349–1354.

[5] Sushree Das, Ranjan Kumar Behara, Mukesh Kumar,Santanu Kumar Rath,Real –"Time Sentiment Analysis of Twitter Streaming data for Stock Prediction". Procidia Computer Science Volume 132,2018.pp.956-964.

[6] Yaojun Wang, Yaoqing Wang Using Social Media Mining Technology to Assist in Price Prediction of Stock Market, IEEE International Conference on Big Data Analysis (ICBDA) March 2016.

[7] Haransha M,Gopalakrishnan E. A.Vijay Krishna Menon,Soman K.P, **"**NSE Stock Market Prediction Using Deep-Learning Models NSE Stock Market Prediction Using Deep-Learning Models". Procidia Computer Science 132(2018).pp. 1351-1362.

[8] Michal Skuza, Andrzej Romanowski "Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction". Federated Conference on Computer Science and Information Systems 2015 pp. 1349–1354.

[9] S. Naveen Balaji, P. Victer Paul, R. Saravanan "Survey on Sentiment Analysis based Stock Prediction using Big data Analytics" International Conference on Innovations in Power and Advanced Computing Technologies 2017.pp.1-5.

[10] G.Shyamala, N.Pooranam, "A Survey on Online Stock forum using Subspace Clustering" 2016 International Conference on Computer Communication and Informatics (ICCCI -2016), Jan. 07 – 09, 2016, Coimbatore,INDIA.

[11] Girija V Attigeri , Manohara Pai M M, Radhika M Pai, Aparna Nayak "Stock Market Prediction: A Big Data Approach". 2015 IEEE Fifth International Conference on Big Data and Cloud Computing Year: 2015.pp.93-98.

[12] Oussama Lachiheb, Mohamed Salah Gouider, "A hierarchical Deep neural network design for stock returns prediction", International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia Procedia Computer Science 126 (2018) 264–272

[13] Paul D. Yoo, Maria H. Kim, Tony Jan, "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation" Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)