

Big Data Analytics for Price Prediction

Kyi Lai Lai Khine, Thi Thi Soe Nyunt
University of Computer Studies, Yangon
kyilailai67@gmail.com, software.ucsy@gmail.com

Abstract

Big Data Predictive Analytics is influenced in the financial market mainly in stock exchange with its emerging technologies. Stock Market Prediction has always been one of the hottest topics in research, as well as a great challenge due to its complex and volatile nature. Stock or share prices are considered to be very dynamic and quick changes because of the underlying nature of financial domain. Therefore, there is a critical need in prediction approaches to be effective and efficient utilization of large amount of market data (Big Data) to analyze future prediction in stock price movement. In this paper, a hybrid prediction model is proposed for predicting daily basis stock price changes or movements. It is based on the combination of historical stock price data and text mining techniques which take the textual contents of Financial News Websites that have highly impacts on price movement.

Keywords: *Big Data, predictive analytics, prediction model, stock price movement*

1. Introduction

Nowadays the Internet represents a big space where great amounts of information are added every day. The IBM Big Data Flood Infographic shows that 2.7 Zettabytes of data exist in the digital universe today. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. The term big data is derived from the fact that the datasets are so large that typical database systems are not able to store and analyze the datasets. The datasets are large because the data is no longer traditional structured data, but data from many new sources, including e-mail, social media, and Internet-accessible sensors. Big data moves around 7 Vs- volume, velocity, variety, value and veracity, variability and visibility. Storing huge volume of data available in various formats which is increasing with high velocity to gain values out it is itself a big deal.

The biggest challenges of big data are facing the seven V's which is:

Volume: The most visible aspect of big data referring to the fact that the amount of generated data has increased tremendously the past years.

Velocity: Capturing the growing data production rates. More and more data are produced and must be collected in shorter time frames.

Variety: The multiplication of data sources where comes the explosion of data formats, ranging from structured text to free text.

Value: The importance or usefulness of the data to those consuming it – is probably the most relevant to organizations.

Veracity: The need to deal with imprecise and uncertain data is another facet of big data.

Variability: There are changes in the structure of the data and how users want to interpret that data.

Visibility: The state of being able to see or be seen is implied. Data from disparate sources need to be stitched together where they are visible to the technology stack making up Big Data.

Big data analytics provides a real opportunity for enterprises to transform themselves into an innovative organization that can plan predict and grow markets and services driving towards higher revenue [2]. Due to chaotic behavior of the stock or share market, traditional techniques are insufficient to cover all the possible relation of the stock price fluctuations. Many researches were the analysis of only using numerical information and the number of proposed methods in financial time series prediction is rely heavily on using historical structured and numerical datasets. But another areas in stock market prediction comes from textual data, based on the assumption that the course of a stock price can be predicted much well by looking at current appeared news articles. Therefore, we want to propose predictive analytics on stock prices with the combination of numerical stock price data and current news from financial websites that have high impact on stock price movement.

2. Related Work

Rama Bharath Kumar, Bangari Shraavan Kumar and Chandrag-iri Shiva Sai Prasad proposed to predict the classified the Financial News based on the contents of relevant news articles which can be accomplished by building a prediction model which is able to classify the news as either rise or drop. They also exploited textual information especially in addition to numeric time series data increases the quality of the input and improved predictions are expected

than only numerical data [6]. Rupinder kaur and Ms.Vidhu Kiran focused on analyzing the historical data available on stocks with accuracy using time series prediction technique in order to help investors to know when to buy new stocks or to sell their stocks. Also, they proposed to develop for a Time-Series Neural Network that achieved a highest percent probability of predicting a market rise and market drop as compare to existing methods [7]. Yuzheng Zhai, Arthur Hsu, and Saman K Halgamuge stated that combining the information from both related news releases and technical indicators is enhanced the predictability of the daily stock price trends. News are trained and classified using SVM and their results feed into another SVM to produce the combined prediction of price trends. They also showed the performance of this system can achieve higher accuracy and return than a single source system [8]. Shital N. Dange, Rajesh V.Argiddi and S.S.Apte presented a new model for future predicting the market direction more accurately when stocks data and textual data are correlated to each other. Feature extraction in this prediction model is used for textual data including parsing of news articles and creation of term dictionary. Decision Tree Induction J48 algorithm is used to generate prediction model for making future prediction and making market action recommendation. The results that produced using this model are more accurate and the accuracy of results hits 84% [1]. J.Kranti and S.S.Apte applied three layered Neural Network in their prediction system. The numerical representation of the stock quotes and the key phrases from news articles are taken as input units to the input layer of multilayered feed-forward network. They also attempted to determine whether the BSE market news in combination with the historical quotes can efficiently help in the calculation of the BSE closing index for a given trading day [5].

Nitish Srivastava stated that dropout is a technique for improving neural networks by reducing overfitting. The main idea is to prevent co-adaptation of hidden units. Dropout improves performance of neural nets in a wide variety of application domains including object classification, digit recognition, speech recognition, document classification and analysis of bio-medical data [11].

3. Theory Background

This section gives some theoretical background of the paper. Firstly, big data analytics, predictive big data analytics and its processing steps are presented. After that, predictive modeling, proposed system's problem statement, design and its workflow are discussed.

3.1. Big Data Analytics

Big data analytics can be defined as the combination of traditional analytics and data mining techniques along with large volumes of data to create a fundamental platform to analyze, model and predict the behavior of customers, markets, products, services and the competition, thereby enabling an outcome-based strategy precisely tailored to meet the needs of the enterprise for the market and customer segment. Analyzing big data is done using Hadoop and it has been widely embraced for its ability to economically store and analyze large data sets. Using parallel computing techniques like MapReduce, Hadoop can reduce long computation times to hours or minutes.

There are three types of big data analytics:

Descriptive Analytics, which use data aggregation and data mining techniques to provide insight into the past and answer: "What has happened?"

Predictive Analytics, which use statistical models and forecasts techniques to understand

the future and answer: "What could happen in future?"

Prescriptive Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer: "What should we do to happen in future?"

Predictive Big Data Analytics comprises a variety of techniques that predict future outcomes based on historical and current data. In practice, predictive analytics can be applied to almost all disciplines– from predicting the failure of jet engines based on the stream of data from several thousand sensors, to predicting customers' next moves based on what they buy, when they buy, and even what they say on social media [4]. Predictive Analytics turns uncertainty about the future into a usable probability. It is also a continuous process. To maximize the success with predictive analytics, following steps must be followed by an organization:

Identify business goals: First step is to clearly identify business goals. Clearly defined business goal can only leads to a successful predictive model.

Data understanding from various sources: After deciding business goal, next step is to collect data from variety of sources available.

Data preparation: Raw data can be collected from variety of sources but preparing that data for predictive analysis is the key challenge. Raw data is unsuitable for analysis. Data preprocessing must be required for run predictive algorithms on the data.

Development of predictive model: Predictive analytics modeling tools are used to run analysis algorithms for the data. Data analysts use one or more of these tools to perform analysis. Hundreds of machine learning algorithms and statistical algorithms are used by data analysts to find predictive models.

Evaluation of the model: Predictive analytics is all about probability not absolutes. Before performing analysis on the test data, organizations used to set a probabilistic output that they will use to compare with the results of the predictive models.

Deployment: Once an effective predictive model is identified, then it is deployed in the production application by the analysts. This

deployed model consists of logic to run predictive rules, formulas, and method to get the data required by the model and finally to obtain the results.

Examine the effectiveness of model and result analysis: It is necessary to continuously evaluate the effectiveness of the model [10].

3.2. Predictive Modeling

Predictive analytics benefits any decision by providing executives, managers and other decision-makers with the tools to make the best possible decision. Modeling is at the heart of predictive analytics and predictive models are born whenever data is used to train a predictive modeling technique. To put it formally, Data + Predictive Modeling Technique = Prediction Model. A predictive model is then the result of combining data and mathematics where learning can be translated into the creation of a mapping function between a set of input data fields and a response or target variable. Many predictive modeling techniques, including neural networks (NNs), clustering, support vector machines (SVMs), and association rules, exist to help translate this data into insight and value. They do that by learning patterns hidden in large volumes of historical data. When learning is completed, the result is a predictive model. After a model is validated, it is able to generalize the knowledge it learned and apply that to a new situation.

4. Proposed System Architecture

4.1. Problem Statement

Due to chaotic behavior of the stock or share market, traditional techniques are insufficient to cover all the possible relation of the stock price fluctuations. Many researches were the analysis of only using numerical information and the number of proposed methods in financial time series prediction is rely heavily on using historical structured and numerical datasets. But

another areas in stock market prediction comes from textual data, based on the assumption that the course of a stock price can be predicted much well by looking at current appeared news articles. News contents are one of the most important factors that have influence on market. Considering the news impact in analyzing the stock market behavior, leads to more precise predictions and as a result more profitable trades. Therefore, we intend to propose predictive analytics on stock prices with the combination of numerical stock price data and current news from financial websites that have high impacts on stock price movement. Big data architecture for stock market is divided into three parts. First job is to identify different data and its sources required for the future prediction of the market. Store, acquire and process of heterogeneity, unstructured and temporal data is the second major challenge. Last part of the architecture is to attain its goal.

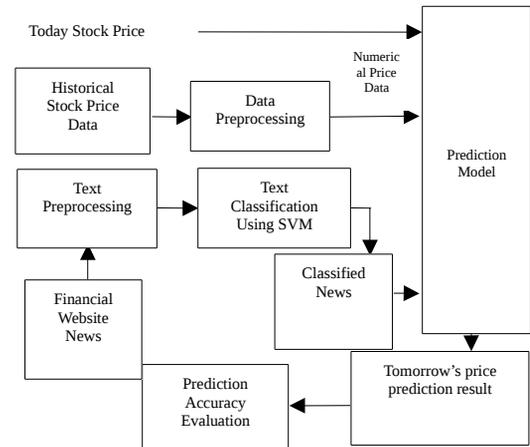


Figure 1. Proposed System's Design

4.2. Proposed System's Workflow

Our proposed system consists of the following major steps:

1. Data Collection
2. Pre-processing
3. Text Classification
4. Predictive Modeling
5. Prediction Accuracy Evaluation

Data Collection is the first step which includes numeric data as well as textual data. Numeric data is history of stock prices, while its related textual data is news articles concerning with stock market and its current price fluctuations condition from financial websites. These financial websites are Financial Times (www.ft.com), Reuters (www.investools.com), Yahoo Finance (www.yahoo.finance.com) and etc. The textual data is collected once a day, and today's closing price data for prediction. Pre-processing is the second step when data has been collected, it is necessary to pre-process both numerical stock prices and unstructured news. In pre-processing unstructured news, stop words removal and stemming procedures have to be done.

I. Stop Words Removal: Input to feed this is .txt file containing stop words list like articles, conjunctions, prepositions etc. as well the URL of any financial news website.

II. Stemming: To fetch the exact grammatical root format of word the stemming process is used. It trims the words e.g. buying to be reduced to 'buy' as root.

Pre-processed news articles are generating keyword phrases corresponding to the given .txt file global list of topmost keyword phrases. These key phrases are initiated with some weight. KEA algorithm is used to obtain single global list of keywords phrases. After having pre-processed both numeric data and textual data, the next step is Text Classification. In

Text Classification, the goal of processing the news generally is to classify the news into two classes either good news or bad news regarding the selected stock. Therefore, we intend to apply Support Vector Machine (SVM) Machine Learning Algorithm. SVM is a popular and highly accurate machine learning method for classification problems. SVM try to find an optimal hyperplane within the input space so as to correctly classify the binary (or multi-class) classification problem.

In Predictive Modeling, historical stock price data, classified news textual data and today's stock price (especially closing price) are used as inputs for prediction model. Neural network based methodologies are the best suited for the stock market forecasting. Therefore, the prediction model is built using trained multilayered feed forward neural network. Backpropagation is the learning algorithm for neural network. A neural network is the set of connected input/output units in which each connection has weight associated with it. During learning phase, the network learns by adjusting the weights so as to able to predict the correct class label of the input tuples. The next day's stock price movement (up, down, unchanged or stable) prediction result are calculated and presented to users. Finally, prediction accuracy measurement is considered as the prediction model evaluation. It is evaluated as the percentage of predictions that were correctly determined by the system.

5. Dropout Technique in Neural Network

One of the problems that occur during neural network training is called Overfitting. The network has memorized the training examples, but it has not learned to generalize to new situations. Large networks are also slow to use, making it difficult to deal with overfitting by combining many different large neural nets at test time. Dropout is a technique for addressing overfitting issue in neural networks by preventing complex co-adaptations on training data. It is a very efficient way of performing model with neural networks. The key idea is to randomly drop units (along with their connections) from a neural network during training. Therefore, we intend to contribute dropout technique in neural network for this proposed price prediction system.

6. Conclusion and Future Work

Forecasting stock prices has always been challenging task for many business analyst and researchers. In fact, stock or share market price prediction is an interesting area of research for investors. Therefore, we try to make use of not only numerical data but also textual data to predict the stock price movement. If we combine both techniques of textual mining and numerical time series analysis, the accuracy in prediction can be achieved. Effective prediction systems indirectly help traders by providing supportive information such as the future market direction. In future work, we will implement the big data processing framework for the proposed system and the effects of dropout technique in neural network for prediction process. And we will consider to apply human decision in prediction model to provide better prediction results for our system.

References

- [1] A. Rajesh, D. Shital, and S. S. Apte, "Financial Trading System using Combination of Textual and Numerical Data", International Journal of Computer Applications (0975 – 8887) Volume 51– No.1, August 2012
- [2] C. Florina, G. Elena, "Perspectives on Big Data and Big Data Analytics", 2013
- [3] C. Hsinchun and P. Robert, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System", 2013
- [4] G. Amir, H. Murtaza, "Beyond the hype: Big data concepts, methods and analytics", International Journal of Management, 2014
- [5] J. Kranti, S. S. Apte, "Stock Market Prediction Model by Combining Numeric and News Textual Mining", International Journal of Computer Applications, 2012
- [6] K. Rama Bharath, K. Bangari Shraavan, "Financial News Classification using SVM", International Journal of Scientific and Research Publications, Volume 2, Issue 3, March 2012
- [7] K. Rupinder, K. Vidhu, "Time Series based Accuracy Stock Market Forecasting using Artificial Neural Network", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015
- [8] K. H. Saman, H. Arthur and Z. Yuzheng, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction", 2013
- [9] M. Suresh babu, N. Geethanjali and B. Sathyanarayana, "Forecasting of Indian Stock Market Index Using Data Mining & Artificial Neural Network", International journal of advance engineering & application, 2011.
- [10] O. James, "The Concept of Predictive Analytics", International Journal of Knowledge, Innovation and Entrepreneurship, 2014
- [11] S. Nitish, "Improving Neural Networks with Dropout", 2013
- [12] W. Huang, Y. Nakamori, and S. Wang, "Forecasting stock market movement direction with support vector machine", Computers & Operations Research, 2005