Using Text Mining and Rule Based Technique for Prediction of Stock Market Price

Jageshwer Shriwas¹, Shagufta Farzana²

¹NRI Institute of Science and Technology, Bhopal (MP)
²Dr. C.V. Raman Institute of Science and Technology, Kota, Bilaspur (CG)

Abstract— Decision making for investor in stock market is considered to be one of the difficult task, because, nature of stock market is dynamic. Prediction of Stock price is the way of trading indexes with the help of key techniques. These techniques are helpful to all the people who want to know how the stock prices are fluctuate. News articles are one of the most important factors which influence the stock market. This paper basically deals with analyzing index point of NSE with the help of using classification rules of data mining to find out the pattern and finally this pattern evaluated with the help of graph based analysis based on daily stock price. This pattern shows new indexes and key point which is average value of all the indexes. And also analyzing the news impact on stock price prediction. Here we are using one year index point of an NSE and analyze it for stock price prediction. This study basically shows the effect of financial news to the prediction of stock market prices as well as daily direction of change in the index.

Keywords— financial news, prediction, ruled based classification, stock price, Text mining.

I. INTRODUCTION

Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large database. Prediction of stock market returns is an important issue in finance. Now a days it is one of the measure role for trading the stock markets. Mining news articles and the time series data concurrently, for predicting the stock market prices is an emerging topic in data mining and text mining communities. Basically prediction is accomplish with the help of news regarding day by day changed and company bench mark. Bench mark of company is based its profit during its production and market gross income. These are come under short term investment of the money in the company. One can find these index value for there better prediction and predict it for future use.

NSE (National Stock Exchange) is one of the leading Stock exchange in India. NSE was established in 1994[15]. The NSE has played a leading role as a change agent in transforming the Indian stock market to its present form. Since its inception, the NSE has been playing the role of a catalytic agent in reforming the stock market and evolving the best market practices.

The NSE has brought about unparalleled transparency, speed and efficiency, safety and market integrity. In this process the NSE has become the largest stock exchange in the country, relegating the Bombay Stock Exchange to the second place. Here we are taking stock market prices and analysis from July 2012 to June 2013 for prediction.

Financial stock market is a complex, non inactive, noisy, disordered, nonlinear and dynamic system but it does not follow random walk process[7,8]. There are many factors that may cause the rise and fall of financial market movement. Predictions of stock market price and its direction are quite difficult. Data mining techniques have been profitably have to shown to generate high prediction accuracy of stock price movement[9,10]. Recently some of the researchers have found that news are one the most influential sources that affect stock market and are necessary in achieving to more accurate predictions[20]. The prediction of stock market is without doubt an interesting task. There are number of methods applied to accomplish this task. These methods use various approaches, ranging from highly informal ways to more formal ways. These techniques are: technical analysis method, fundamental analysis method, Traditional time series methods and machine learning methods[6].

Technical analysis is an alternative approach to the study of stock price behavior. The rational behind technical analysis is that share price behavior repeats itself over time and analysis attempt to derive methods to predict this repetition. A technical analyst looks at the past share price data to see if he can establish any patterns. he then looks at current price data to see if any of the established patterns are applicable and, if so, extra potations can be make to predict the future price movements[17]. Technical analysts, known as chartists, attempt to predict the market by tracing patterns that come from the study of charts which describe historic data of the market. Charting represents a key activity in technical analysis, because graphical representation is the very basis of technical analysis. It is a security prices that are charted.

Data mining refers to the process of finding interesting patterns in data that are not explicitly part of the data[10,11].
Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k means clustering is unsupervised) [13]. Classification techniques are a type of supervised learning of data mining. Classification is define by the set of predefined classes from a training set as a result of learning from that dataset [14]. News contents are one of the important factor to effect the stock market. A successful news analysis would be achieved if the effective information about stock could be extracted from the news article. It is must to extract the key phrases which may effect the stock price because we know that news articles are unstructural or textual information. Now question is arise how the news impact can be used in stock Price prediction. Basically automatic text classification techniques are used to analyze the incoming news.

However, this paper is focus on, to improve the accuracy of stock market price prediction by using combine approach of rule based classification and technical analysis as well as effect of financial news also. The rest of the paper consist following sections as followed. Next proposes impact of financial news in stock price prediction. Next Section introduces the basic concept of classification rules and its uses for prediction the stock price. Next section shows the analytical relationship between index points. And last section conclude the paper.

II. REVIEW OF LITERATURE

Anil Rajput et.al.[10] focuses for the rule based classification model of historical BSE stock data with data mining techniques. They used decision tree and rule induction method with the help of data mining software. They have to create classification rules and induction rules with the use of J48 and PRISM classifiers under WEKA software. Stephen Evans[4] uses classification data mining to attempt to predict the direction of daily returns of randomly selected stocks from the Russell 1000 and Russell 2000 stock indexes. The study uses moving averages of historical daily stock prices as attributes, along with different data mining classifiers, to attempt to make these predictions. A secondary goal of this study is to determine how effective using Distributed Data Mining (DDM) can be in predicting the direction of daily stock returns. Hence, DDM classifiers are used in the testing. This study discovers that the moving averages of daily returns do not help predict the direction of future daily stock returns any better than the percentages of returns from one trading day to the next. It also shows that the classifiers were no more than 60% accurate in predicting the directions of daily returns for any of the stocks used in this study.

Hence, it appears that momentum cannot be used to explain very much of the movement in daily stock prices on a consistent basis.

S. Karthik et.al.[15] objective is to analyze the stock market trend using integrated clustering and weighted rule mining technique. For predicting and analyzing the market trends data mining techniques are also used. Statistical techniques are used for the market price prediction process. Inaccurate results are produced in the statistical analysis. To find the stock market trend with index dependency analysis environment hybrid clustering and association algorithm is not appropriate. Statistic analysis techniques are not suitable for trend analysis with index relationship. The stock market transactions data is analyzed with clustering and weighted rule mining techniques. The K-mean clustering algorithm is used to cluster the transaction with respect to the market flows. The market trade transactions are divided into three zones such as up trend, down trend and stable zone. The weighted rule mining technique is applied to fetch patterns from the indexes, sector indexes and company price values. Apriority algorithm is modified to carry out weighted rule mining process. The system produces the market trade trend flow with market indexes and sector index values.

S. Prasanna et al.[18] have discussed several attempts made by researches for stock price prediction. These works show that data mining techniques can be applied for evaluation of past stock prices and acquire valuable information by estimating suitable financial indicators. They discussed works basically two types of prediction methods are implemented by several researches to generate useful extracts. They are fundamental approaches and technical indicator based approaches. Many researchers adopted technical indicator approaches only. Limited work is done with fundamental approaches which give plenty of opportunity for further research. Phichhang Ou et. al[9],discussed ten different techniques of data mining and applied to predict price movement of Hang Seng index of Hong Kong stock market. The approaches include Linear discriminate analysis (LDA), Quadratic discriminate analysis (QDA), K-nearest neighbor classification, Naive Bayes based on kernel estimation, Logit model, Tree based classification, neural network, Bayesian classification with Gaussian process, Support vector machine (SVM) and Least squares support vector machine (LS-SVM).Experimental results show that the SVM and LS-SVM generate superior predictive performances among the other models. Specifically, SVM is better than LS-SVM for in-sample prediction but LS-SVM is, in turn, better than the SVM for the out-of-sample forecasts in term of hit rate and error rate criteria.
Manisha V. Pinto et al. [19] provide a framework for predicting stock magnitude and trend for making trading decisions by making use of a combination of Data Mining and Text Mining methods. The prediction model predicts the stock market closing price for a given trading day ‘D’, by analysing the information rich unstructured news articles along with the historical stock quotes. In particular, we investigate the immediate impact of the news articles on the time series based on Efficient Market Hypothesis (EMH). Key phrases provide semantic metadata that summarize and characterize documents. This framework incorporates Kea [21], an algorithm for automatically extracting key phrases from news articles. The prediction power of the Neural Network is used for predicting the closing price for a given trading day.

The Neural Network is trained on the extracted key phrases and the stock quotes using the Back propagation Algorithm. Nowadays, in stead of a single method, traders need to use various prediction techniques parallel to gain multiple signals and more information about the future of the markets.

### III. FINANCIAL NEWS AND ITS ROLES FOR PREDICTION

Financial News based stock market prediction can be considered as a text classification task. Generally the goal is to forecast some aspects of the stock market such as price or volatility based on the news content. Information in the form of quotes and financial news is released into the market all the time. While quotes data is structured and can be directly used, the challenge is to manage the large amounts of textual information. We can employ techniques to parse the news articles and identify the key features most likely to have an impact on the stock market. By automating this process, machines can take advantage of arbitrage opportunities faster than human counterparts by repeatedly forecasting price fluctuations and executing immediate trades to make profits [19]. At a time lots of news articles are released. We know that news articles are an unstructured form of data with enormous amount of useful information embedded in it. KEA [21], an algorithm for automatically extracting key phrases from text is used [19,21]. The most common key phrases extract from the news articles which influence the stock prices are listed below:

<table>
<thead>
<tr>
<th>Table 1: Shows top 20 key phrases [19]</th>
</tr>
</thead>
<tbody>
<tr>
<td>'opens lower'</td>
</tr>
<tr>
<td>'jobless claims'</td>
</tr>
<tr>
<td>'open higher'</td>
</tr>
<tr>
<td>'slumping dollar'</td>
</tr>
<tr>
<td>'battered financial'</td>
</tr>
<tr>
<td>'oil futures closing high'</td>
</tr>
<tr>
<td>'unemployment rate'</td>
</tr>
<tr>
<td>'oil prices'</td>
</tr>
<tr>
<td>'lifted financial shares'</td>
</tr>
<tr>
<td>'credit crisis'</td>
</tr>
<tr>
<td>'strong gains'</td>
</tr>
<tr>
<td>'jobless claims data'</td>
</tr>
<tr>
<td>'drop in crude oil futures'</td>
</tr>
<tr>
<td>'economic news'</td>
</tr>
<tr>
<td>'investors' confidence'</td>
</tr>
<tr>
<td>'quarterly loss'</td>
</tr>
<tr>
<td>'nuclear power'</td>
</tr>
<tr>
<td>'investors' appetite'</td>
</tr>
<tr>
<td>'nuclear crisis'</td>
</tr>
<tr>
<td>'required reserves'</td>
</tr>
</tbody>
</table>

Based on prediction goal a set of final classes are defined, such as “Up” (which means this news cause the prices to go up), “Down” (which means this piece of news is probable to causes decrease in prices) and etc. The prediction system is supposed to classify the incoming news into one of these classes. News based market prediction can be divided into two main phases. “Training phase” and “Operational phase”. In operational phase, one of the predefined classes will be assigned to incoming news; however, to make the system ready for the operational phase a classifier should be trained in the training phase. Machine learning techniques are widely used to automate such processes. As a part of the training phase, a set of training data shall be prepared which in our case the train data are the pre-classified news and market information such as market prices. These labeled (pre-classified) news and possibly market numerical data will be processed to be fed into the classifier for training. The trained classifier would be ready to get a piece of news and assign a class to it in operational phase.
The accuracy of the system is measured as the percentage of the predictions that were correctly determined by the system. For instance, if the system predicts an uptrend and the index indeed goes up, it is assumed to be correct, otherwise, if the index goes down or remains steady for an uptrend, it is assumed to be wrong.[19,21]. Representing the news section consists of two main tasks: feature selection and feature weighting. First, a set of features will be selected to represent a piece of news and next step is to assign weights to those selected features. These weighted vectors would be the inputs to classifier. In real-world trading applications, the amount of textual data available to stock market traders is staggering. This data can come in the form of required shareholder reports, government-mandated forms, or news articles concerning a company’s outlook. Reports of an unexpected nature can lead to wildly significant changes in the price of a security.

IV. CLASSIFICATION RULES AND THEIR APPLICATION IN PREDICTION

Classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. The models are used to predict the class of objects whose class label is unknown. A bank, for example, may classify a loan application as either a fraud or a potential business using models based on characteristics of the applicant. A large number of classification models have been developed for predicting future trends of stock market indices and foreign exchange rates[3]. The objective of classification is to first analyze the training data and develop an accurate description or a model for each class using the features available in the data. Such class descriptions are then used to classify future test data in the database or to develop a better description (called classification rules) for each class in the database[16]. Classification is perhaps most familiar and most popular data mining technique. Classification application include image and pattern reorganization, medical diagnosis, loan approval detecting faults in industry application and classifying financial market trends. Estimation and prediction may be viewed as prediction a continuous values, while classification forecast a discrete values. Classification are categories in many groups. These are Statistical based, Distance based, Decision tree, NN and Rule based. Rule based classification are used to define if-then rules that can be used to explain the result of given training sets[4,5].

A. Rule Based Classification

Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of if-then rules for classification. Rule-based classification is a type of supervised learning data mining method[10]. A key approach to generate subsets from given data set through if-then rules that cover all the appropriate conditions. Expression of if-then rule is:

If condition then conclusion

For example:-

R1: If age=youth AND student=yes then buys-comp=yes

The if part (or left side) of a rule is known as the rule antecedent or precondition. The then part (or right side) is the rule consequent. In the rule antecedent, the condition consists of one or more attribute tests (e.g., age and student) that are logically ANDED. The rules consequent contains a class prediction (in their case, we are predicting whether customer will buy computer buys-comp is yes). R1 also written as:

R1: (age=youth) \& (student=yes) \rightarrow (buys-comp=yes)

If the condition (i.e., all the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply that the rule is satisfied) and that the rule covers the tuple. A rule R can be assessed by its coverage and accuracy. Given tuple X, from a class-labeled data set D, let n_cover be the number of tuples covered by R; n_correct be the number of tuples correctly classified by R; and |D| be the number of tuples in D. We can define:

Coverage (R) = \frac{n_{correct}}{|D|}

Accuracy (R) = \frac{n_{correct}}{n_{cover}}

That is, a rule’s coverage is the percentage of tuples that are covered by the rule (i.e., their attribute values hold true for the rules antecedent). For a rules accuracy, we look at the tuples that it covers and see what percentage of them the rule can correctly classify.

B. Use of Classification rules in this paper

This paper is based on data sets analysis of one year index point of NSE. We are categorized all the data sets into subsets of its period of month by the use of classification rules. This will help to find an average value for individual months. Now this individual average values also categorized in such a way that one can examine its final value, this is our classified value which is known as a pattern. The more remarkable use of this value in prediction because we can distinguish all the data sets in single set for pattern.
Technical analysis is the method of predicting the share prices move in trends that are based on the constantly changing attributes of investor. Using technical data such as price, volume, highest and lowest prices per trading period, technical analysts believe that share prices are determined by the demand and supply forces operating in the market. These demand and supply forces in turn are influenced by a number of fundamental forces as well as certain psychological or emotional factors. Many of these factors cannot be quantified. The combined impact of all these factors is reflected in the share price movement. A technical analyst therefore concentrates on the movement of share prices. He claims that by examining past share price movements future share prices can be accurately predicted. The basic premise of technical analysis is that prices move in trends or waves which may be upward or downward. It is believed that the present trends are influenced by the past trends and that the projection of future trends is possible by an analysis of past price trends. A technical analyst, therefore, the price and volume movements of individual securities as well as market index. Thus, technical analysis is really a study of past or historical price and volume movements so as to predict the future stock price behavior[17]. The technical analyst uses charts to predict future stock movement. Price charts are used to detect the trends or daily movement of market. These trends and movements are assumed to be based on supply and demand issues which often have cyclical or noticeable patterns. Through graphical analysis we calculate the average index value of each month and assigned in AVG variable. Then after we calculate the final pattern or final average from the above average values and this final pattern is assigning in FAVG variable. Final average is nothing it is average of AVG values. Once we got the final pattern then we apply if-then rules to predict the price of stock market. To achieve more accuracy we apply if-then rules in three steps. These steps are explained below:

**Step 1:**

// Compare the data of each day to AVG

If AVG=D1 then count in value1

(D1 is daily index value)

**Step 2:**

// Compare daily index value to FAVG

If D1=FAVG then count in value2

**Step 3:**

// Compare the data of value1 and value2

After comparing the values i.e., value1 and value2. The number which repeats or which exists for maximum time will be taken as final value or predicted index value. Parallely we compare daily stock price to daily release news articles. News content is unstructured form so we used the key phrases extracted from news article mentioned in table 1. If coming news having key phrases which listed in table 1 then news go with up class means increase the price of stock otherwise decrease the stock price. This will help the investor to predict the stock price or provide a way to invest which gives more profit.

**VI. CONCLUSION**

This paper presents the set of classification rules of NSE to predict the stock market trend. These rules are generated by combining approach of graphical analysis and rule based classification of data mining. Simultaneously we check the effect of news article to daily stock price also. First we found average pattern from graph based analysis. This is calculated as to take the average of each months according daily stock price. After that all patterns are classified according to there time period now we find a final pattern FAVG. Then we compare the data of value1 and value2 which value comes maximum times this value is choose as a final predicted value. So as per our approach, we came to conclusion that our approach is time consuming as we have to compare number of data’s at a time which is practically not possible to do prediction but with point of view of studying we can achieve the main objective with accuracy.

So, for the future research we require more efficient way to achieve the goal with better efficiency and accuracy both.
REFERENCES

[2] Jiawei Han, M. Kamber, J. Pei, "Data Mining Concepts and Techniques." Impression, 2012.
[16] Ming-Syan Chen, Jiawei Han, Philips S. Yu, "Data Mining: An overview from a Database Perspective" IEEE Transactions on Knowledge and Data Engineering, Vol 8, No 6 December 1996.