STOCK MARKET FUNDAMENTAL ANALYSIS USING DATA MINING

1YUVARAJ A.N, 2CHANDRIKA J

1Student, Malnad College of Engineering, Hassan
2Professor, Malnad College of Engineering, Hassan
E-mail: 1yuvaraju.an@gmail.com, 2jc@mcehassan.ac.in

Abstract— Stock market plays a vital role in a country’s economy and it is an important consideration in all the fields due to its potential financial gain. Using Data mining future stock price, whether stock price go high or low can be predicted. There are two types of stock market predictions, fundamental analysis based on unstructured data and technical analysis based on structured data. In this paper different between these two are presented and also a methodology to fundamental analysis has been proposed.

Keywords: Data Mining, Fundamental Analysis, Stock Market, Sentiment Analysis, Text Mining.

I. INTRODUCTION

The stock market prediction using data mining technology is one of the most important issues to be investigated. In this paper different methods of stock market analysis using data mining and Text mining that predicts the changes of stock price by analyzing structured data and unstructured data respectively has been discussed.

1.1 Fundamental analysis:
In fundamental analysis the stock price will be predicted by examining related financial, economical and other factors. In fundamental analysis it considers all factors including macro elements that can affect the stock price. The value produced after performing the fundamental analysis is compared with the current market price so that the investor configure out what position to take with respect to stock, if the analysis is under price = Sell, overpriced = Buy. Fundamental analysis is based on unstructured data like news, articles, xml files, email. The Fundamental analysis is considered to be the opposite of technical analysis.

1.2 Technical analysis:
In technical analysis the stock price will be predicted by considering the past stock movements which is similar to weather forecasting, this analysis may not give accurate result about the future predictions, but allows the investor to anticipate what likely to happen during period of time.

Technical analysis uses wide variety of charts like doji, double top, cloud formation etc. Technical analysis makes use of the structured data, in this everyday’s open, close, high, low and volumes of stock traded are captured and stored in the database.

II. LITERATURE SURVEY

In the paper “A Data mining algorithm to analyze stock market data using lagged correlation” [2], authors Cecil Fonseka and Liwan Liyanage have explained several ways for the investor to reduce risk in short term investing, using lagged correlation algorithm for predicting the stock price direction when two stocks are strongly correlated for many number of days. It follows the technical analysis method. In the paper “Predicting stock prices using data mining techniques”, [8] authors Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi gives an idea of predicting stock market using the historical data, which helps the investor to take the right decision on their stock. This method is technical analysis which is not perfect because it won’t include many factors. Authors Lipika Dey, Anuj Mahajan and SK. Mirajul Haque in the paper , “Document Clustering for Event Identification and Trend Analysis in Market News”.[1] presents text clustering method for financial news, which is significantly differing from the previous works of Stock market analysis.

In this paper of event extraction Latent Dirichlet Allocation method has been used for fundamental analysis. In the paper , “Stock Market Prediction Using Data Mining”, [5], authors Ruchi Desai, 2Prof.Snehal Gandhi paper makes use of the NLP based module & statistical parameter based module which helps to find out the sentence polarity and comparison between previous data so that fundamental analysis can be done.

The authors Rasika Amarasiri1, Jason Ceddia, Daminda Alahakoon in the , “ExploratoryData Mining Lead by Text Mining Using a Novel High Dimensional Clustering Algorithm”, [7] gives the importance of text mining for the unstructured data and also provides high dimensional clustering algorithm for text mining which explores data with text. Same has been implemented the real time data. In the paper “Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing”, authors F. S. Gharehchopogh, Z. A. Khalifelu [9] the methods are proposed for converting the unstructured data to structure data to find patterns in textual unstructured files based on contents and Tried NLP to reach concepts of texts via specific algorithms.
III. KNOWLEDGE DISCOVERY IN TEXT (KDT):

The term KDT is used to indicate the overall process of turning unstructured textual data into high level information and knowledge, while the term Text Mining is used for the step of the KDT process that deals with the extraction of patterns from textual data. By extending the definition of KDD, the following simple definition is given: Knowledge Discovery in Text (KDT) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured textual data.

Text Mining (TM) also known as text data mining is a step in the KDT process consisting of particular data mining and Natural Language Processing (NLP) algorithms that produces a particular enumeration of patterns over a set of unstructured textual data. There are various definitions and terminologies for text mining provided by different researchers. KDT is a multi-step process, which includes all the tasks from gathering of documents to the visualization and evaluation of the extracted information.

3.1 Structured data

Data which resides in fixed format within a file or record are called as structured data. It includes a data in relational database, XML and spread sheets. In structured data the first step is creation of model, the model is type of business data that recorded, stored and how it is accessed. It includes defining what fields are required and how they are stored, what types of data type are used (int, date, name, currency and numeric) and constraints in data input (date within the given period).

Structured data has advantage easy CRUD operation along with the analysis. In ancient days because of technology cost and performance limitations the relational database and spread sheets are the only ways of processing data effectively. The data apart from these two structures are usually stored in the format of text which is an unstructured data.

Structured data is often managed using Structured Query language (SQL) which is used perform CRUD operation in an relational database management systems which is developed by IBM. Structured data has a huge improvement over a period of time in terms of storage and performance. But in real world always data cannot be put in to specific format so structured data is supplemented by the unstructured data and semi structured data.

3.2 Unstructured data

Previously data mining methods for stock market prediction are using the structured data such as relational database, transactional database and data warehouse. However a valuable information is stored in the text format, which contains huge amount of information from different sources such as news, books, articles, web pages, XML Pages email etc. At present days most of the information is stored in text form in all the fields such as Government, industries, business, research centers etc. This text format generally stored in text database which is generally semi structured data such as spread sheets and XML Pages.

Till last decade quick and easy availability of information was not possible, at present news, information is easily accessible, as content provider and content locators are present on the World Wide Web.

IV. INFLUENCE OF NEWS ARTICLES ON STOCK MARKET:

News which is related to market and stock exchange contains important messages of political and economical events. This information plays a vital role in the market predictions. Now a days many financial information available on the Web from giant company’s website like NDTV Profit, CNBC TV18, these websites release the daily issues in electronic versions which contains global and regional political news, economical news and also recommendations from the financial experts.

Many researches had confirmed the reaction of news articles to stock market prediction. They have shown that economic news always has a positive or negative effect on the number of traded stock. They used salient political and economic news as proxy for public information. They have found that both types of news have impact on measures of trading activity including return volatility, price volatility, number of shares traded, and trading frequency.

Fig 1 shows that the stock price goes high because of positive annual budget in the month of march in every year.

![Fig 1: Indication of increase in stock price during annual budget](image-url)
V. METHODOLOGY

It consists of mainly five phases.

- Archive News
- Extract Features
- Feature selection
- Archive Stock Value
- Machine learning algorithm

5.1 Archive News

Nowadays financial news is widely available in the Internet mainly from the websites. Most of the websites have RSS Feed. RSS feed is the main source of information collection. When RSS feed is specified in the project, whenever the websites publish new articles the project will get that article automatically.

5.2 Extract Features

To extract features from unstructured data, we make use of (Part of speech tagger) POS, it is software which reads the data in some language and allots part of speech for each word given unstructured data such as verb, noun, adjectives etc.

Part-of-speech tagging (POS tagging or POST), also called as tagging, where two types of tagging are there manual and automatic. In the process tagging a word in a text corresponds to particular part of speech based on the context and definition.

POS-tagging algorithms can be broadly categorized into rule base and stochastic. Tagging process is harder because some words represents different part of speech at different times and some part of speech are unspoken and complex. In automatic tagging its difficult to make difference from 50 to 150 distinguish part of speech for English language. For example, NP for singular proper nouns, NN for singular common nouns, NNS for plural common nouns.

The POS tagger we are proposed use is the project is developed by Stanford University natural language processing group; it is licensed under the GNU general public license as it is an open source.

5.3 Feature selection

Once the features are extracted from the unstructured text, the interested features are selected. To select the interested features e make use of the SentiWordNet_3.0.0, this is a software resource for opinion mining. It allocates three values for each word called sentiment scores: negative, objective, positive. In resent researchers trying to research on automatic determinism of PN-polarity of the selected terms, in order to identify whether the given word has positive or negative connotation. SentiWordNet is openly available for research works and it is integrated with web based graphical user interfaces.

5.4 Archive Stock Value

In order to predict stock, based on statistical parameter like previous day open and close value, it is essential to collect previous days data in terms of daily open and close price, and the difference between these two values are to be calculated to compare with up down values mentioned in new sentences.

Previous day’s data can be obtained by different sources such as:

- Websites which publish financial news articles: http://economictimes.indiatimes.com/
  http://www.efinancialnews.com/
- Stock related corporate companies: BSE NSE MCX
- From brokerage companies: ShareKhan, Money Control, KARVY, Indiabulls etc.

5.5 Machine learning algorithm

The Machine learning algorithm builds a model from example inputs and making use of that, predicts the decisions. Several machine learning algorithms are available as listed below.

1. C4.5
2. FP-growth algorithm
3. k-Means
4. Support Vector Machines (SVM)
5. Apriori
6. Expectation Maximization (EM)
7. PageRank
8. AdaBoost
9. k-Nearest Neighbors (kNN)
10. Naive Bayes
11. Classification and Regression Tree (CART)

Here we propose the FP-growth algorithm because in this algorithm it needs to scan the database only twice: during first scan it gets frequent 1-itemsets, during the second scan uses 1-itemsets and filters infrequent items and generate FP Tree. This algorithm reduces the cost of searching. This algorithm more scalable, efficient and faster than the other algorithms [31]

CONCLUSIONS

In Data Mining, to predict stock market a methodology has been proposed based on fundamental analysis which considers the unstructured data such as Articles, web pages, financial news etc., The paper also discusses the differences between the structured data and unstructured data and also the advantage of using the unstructured data.

By using the proposed methodology a reliable and fast stock market prediction can be obtained which gives the customer a better option for where to invest the valuable money.

REFERENCES


