International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

STOCK MARKET ANALYSIS USING ONLINE NEWSFEEDS

S.A.Mahajan Assistant Professor Department Of Computer Engg. & Information Tech. PVG's COET; Atharva S. Deshpande, Bachelor of Engineering, Information Technology, PVG's COET; Manasi R. Aminbhavi, Bachelor of Engineering, Information Technology, PVG's COET; Aslesha P. Barke, Bachelor of Engineering, Information Technology, PVG's COET; Rahul Bhagwat, Bachelor of Engineering, Information Technology, PVG's COET.

Abstract

The Stock Market is a volatile and chaotic dynamic system and the reason particular stock rise and fall can be complex. More often than not, stock prices are affected by a number of factors and events some of which influence stock prices directly and other do so indirectly. In the paper, we analyse how online news feeds affect stock market. The system we propose will make use of Natural Language Processing techniques to extract information from online news feeds. Further, the extracted information will be parsed into template format. The system proposes on developing a Decision Support System (DSS) that takes into account market trends, financial analysis and strategies to identify the best time to purchase stocks and what stocks to purchase. The system further visualise the data using googleVis package in R language, which is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing, to help investors manage their portfolios.

Keywords - Natural Language Processing(NLP); Decision Support System(DSS); Parsed; googleVis; R.

Introduction

A stock market is combination of buyers and sellers. The stocks are listed and traded on the stock exchanges, which are the entities of an organization specialized in the business of buying and selling the stocks. The stock prices often fluctuate , so the investors have to be aware of the stock changes taking place. There are such facts and traits about the stock market which makes it extremely difficult for beginners to work in the stock environment[1]. It is with this thought, the Decision Support System (DSS) is proposed. Stock market is depended on many factors. Previous systems were focusing on statistical data for analysis but stock market is also influenced by recent events whose trails cannot be found in statistical data. Recent News cover the recent events ,so we can use these for firm analysis of stock market.

The system proposes on using Natural Language Processing (NLP) techniques to extract information from online news feeds and then use this information to predict changes in stock prices . For example, company names can be recognized and simple template can be filled using parsing of words [2]. These templates can be clustered into groups which can be correlated with changes in the stock prices.

The stock market contains a huge amount of data that vary over time. Finding useful patterns in stock market data requires tremendous analytical skills and effort. To help investors manage their portfolios, we will be developing a system for clustering and visualizing stock market data. Our system intends to assist users in identifying groups of stocks having price movements over a period of time. The System will also be representing the data, in a way, that analyzes the historical price movements of companies, and visualises the data using googleVis, a package in R, which will help user in better understanding of stock market. R is a programming language and software environment for computing and graphical representation.

Existing System

The stock market is a complicated environment. Investors face higher risks to invest in stock market compared to other form of financial investments . Many people have tried to predict the changes of stock prices and predict the market but no one could accurately predict the changes of a particular stock. The existing systems can't be identical to each other. Each one provides a unique price and feature point. Hence, there is a problem whether one system will work with another system or not. Most of the existing systems require input in the form of specific stock data with specific attributes of the stocks. Data is hardly tested on the live market and identification of these attributes is very difficult[1]. Information Technology (IT) professionals are trying to utilise the stock price prediction area using the Artificial Intelligence (AI) technique . There are many approaches in this field, but there is no sign of development of a wide-ranging system. Data Mining, Artificial Neural Network (ANN) and Regression

Analysis techniques are being used to create a wide-ranging system. Various Genetic Algorithms, Time Series Analysis, Fuzzy Neural Networks are also being used to create an effective stock prediction system. A prototype model, Multilevel and Interactive Stock Market Investment System (MISMIS) has also been put forward to forecast stock prices. Clustering techniques are considered effective for analysing stock price. The best-known partitioning clustering algorithm is the K-means algorithm . The algorithm is simple, straightforward and is based on the firm foundation of analysis of variances.

Existing system which was developed focused on statistical data i.e. Historic Prices and Stock Parameters for analysis. Stock market being depended on many factors, solution provided was incomplete and did not focus on all the aspects of the stock market.

Methodology

A. Data Retriever

Our Our system will be using data retriever for crawling information from web pages. Historic Data, Local Stock Parameters and company profiles will be will be retrieved from Yahoo Finance India website, Economic Times & India Infoline[1]. Yahoo Finance websites will also be used for retrieving recent news related to stock market.

Kimono Labs is an online tool which will be used for retrieving the data. It allows us to select the desired data points from the websites. It crawls data and presents it in the form of an API. Kimono returns JSON objects and CSV file which will be used in our system for further analysis.

Number of API will be created for retrieving different kind of data. There will be custom API's for,

i) Retrieving General Stock information.
ii) News feeds from various websites
iii) Local stock parameters.
iv) Historic Stock price

B. Natural Language Processing(NLP)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the communication between computers and human languages. As such, NLP is related to the human–computer interaction. Natural Language Processing (NLP) is a technique that makes pre-processing of input text a smooth process. Pre-processing means cleaning and normalisation of text.

NLP task is divided into two components.
1. Message Understanding Component.
2.Statistical Co-relation Component.

1) Message Understanding Component:

In this, automatic filling of simple templates is done. RJSON, RJSONLITE and RJSONIO are the R packages that will be used for reading JSON objects into R from JSON(). It will further be used for parsing JSON objects. Top down Chart parser will be used for parsing text and converting it into tokens. After getting the tokens, locating and classification of the tokens into predefined categories is done using the Named Entity Recognizers.

Named Entity Recognizer -

Named Entity Recognizer also known as Entity Identification , Entity Chunking , Entity Extraction is a part of information extraction. Extraction attempts to locate and classify elements in the text into predefined categories such as Names of persons, Organizations, Locations, monetary values, percentages etc.

Input to NER will be the tokens generated from parsing process. These tokens will be Nouns, Verbs, Adjectives, Prepositions, etc.NER will classify these tokens according to their category. Output of the NER will be predefined templates.

Ex. TCS price down by 20 %.
Template format for this news feed can be-
Company - TCS
Item - Price
Action - Down
Relative change - 20%

For this Purpose in R we will be using a low level connection with Java using the .rjava package and other packages like 'openNLP' and 'qdap'. With the help of these packages we will further use annotators for the .txt file to mark the sentences, words, names, organizations.

2) Statistical Co-relation Component :

In this, testing of the association between the patterns obtained from message understanding component is done , to determine increases or decreases in the stock prices.

Groups of equivalent Words (e.g., “announced”, “reported”, “released a report” . . .) can be initially determined using online thesauruses such as

Word Net, and then refined using statistical co-occurrence data (e.g.: Words that tend to show up in the same environment belong in the same group). Other important word groups include different actions (hiring, buying, selling . . .) and types and directions of change (increase, decrease, improve, Worsen.) More descriptive words such as "breathtaking,”“shabby,”“askance,""improprieties,”“titan,” etc., may also prove useful.

C. Perceptron Backtracking

In existing system, Perceptron was used to classify output into two classes to predict whether we should buy or sell the stock. In our system Multilayer Perceptron Backtracking will be used to adjust the weights (threshold) based on previous results.

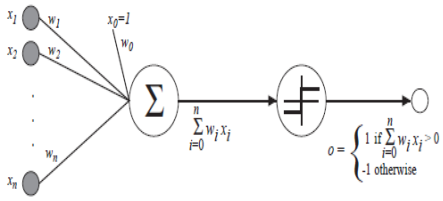

Figure 1: Perceptron Model for DSS

Multilayer perceptron (MLP) is feed forward artificial neural network model that maps sets of input data on to a set of appropriate outputs. A MLP consist of multilayer of nodes in a directed graph with each layer connected to the next one. MLP uses a supervised learning technique called back propagation for training the network. MLP is a modification of standard linear perceptron and can distinguish between data that are not linearly separable.

Perceptron is used to classify input into two classes, whether we should buy or sell that particular stock[6]. Output of all the previous processes will be feed into perceptron. i.e Historic prices, Stock parameters, clustered templates, forecasted prices.

Each input will be assigned weight based on its importance on scale of 0-5. Summation of all the inputs and corresponding weights will feed to perceptron where it will be checked against a predefined threshold value. If the output exceeds the threshold perceptron will hit 1 which signifies that the stock is profitable and we should buy this stock. If output does not exceed the threshold perceptron will hit 0 which signifies that stock is not profitable.

Backtracking algorithm will be used to backtrack through the previous results and learn from the previous results to adjust weights of input parameters and threshold value.

D. Data Visualisation

Data visualization is a technique that describes efforts to help people understand the importance of data by placing it on a visual context. In other words, it is representation of data in pictorial or graphical format.

The stock market contains a huge amount of data that varies over time. The stock price of a company is determined by various factors like the performance of the company , the condition of the economics in general. Fund managers and investors have to analyze stock market data regularly to identify profitable and non-profitable stocks depending on their investment goals and time frame. Finding useful information in such complex data needs high analytical skills and efforts. Thus, for helping users to get a better insight to the stock market our system will be visualising the data using googleVis.

1) googleVis

googleVis is a package in R which provides an interface between R and the Google Chart Tools[3]. The main function of this package is to allow users to visualize data with the Google Chart Tools without uploading their data to Google. The output of googleVis functions is a html code that contains the data and references to JavaScript functions hosted by Google. To view the output a browser with an internet connection is required, the actual chart is seen in the browser eg: Flash.

This package provides interface to Motion Charts, Annotated Time Lines, Maps, Geo Maps, Geo Charts, Intensity Maps, Tables, Gauges, Tree Maps, further Line, Bar, Bubble, Column, Area, Stepped Area, Combo, Scatter, Candlestick, Pie, Sankey, Annotation, Histogram, Timeline, Calendar and Org Charts.

Our system will be performing visualisation in 3 different type:

1)General Market Analysis:

In General Market Analysis, our paper will show a general idea to the user about current situation of the stock market. It will be including the Fast Growing stock, Least Risky stock, Best PEG's etc. This will provide a basic idea to the user of the stocks in the market.

a)Least Risky & Most Risky: This will be used to find the 5 least risky stocks and 5 Most risky stocks. A stock's beta coefficient is used to measure least risky stocks. It is a measure of its volatility over time compared to a market benchmark. A beta of 1 means that a stock's volatility matches up exactly with the markets. A higher beta indicates great risky, and a lower beta indicates less risky. We need to calculate beta manually.

Calculating Beta:

Step1: The stock's closing price for each day for a given period of time and closing level of a market benchmark are taken.

Step 2: Then calculations for the daily price change for both the stock and index as a percentage using the following formula:

$$\text{Daily \% Change} = (\text{Today's Price} - \text{Yesterday's Price}) / \text{Yesterday's Price} \times 100 \quad (1)$$

Step 3: Then comparison between how the stock and the index move relative to each other ie. beta ,using the below formula:

$$\text{Beta} = \text{COVARIANCE}(\text{Stock's \% Daily Change}, \text{Index's \% Daily Change}) / \text{VAR}(\text{Index's \% Daily Change})$$

$$\text{Beta} = \text{COVARIANCE.P}(C2{:}C21, E2{:}E21) / \text{VAR}(E2{:}E21) \quad (2)$$

b) Top & least expensive: This will be used to determine Most expensive and least expensive stocks. PEG ratio of Indian stocks is a very precise value indicator for investors.PEG is the ratio of PE(Price/Earnings) and EPS(Earnings Per Share) growth rate. A combination of PEG and PE becomes more reliable PEG ratio which helps investors estimate true value of stocks. It is highly believed that high PE ratio represents over valued stocks.

Step 1 : Calculate PEG ratio

PEG = PE/EPS

If PEG ratio < 1 means

i) Stock is undervalued
ii) We can expect future growth in market price of stock as its EPS will rise in time to come.
If PEG ratio > 1
i) Stock is over valued
ii) We cannot expect growth in future.
c) Growing Stocks
Step 1 : Calculate EPS for one year.

(Current Year EPS - Prior Year EPS) / Prior Year EPS

This will give one year growth rate. Calculate 3 year, 5 year, 10 year growth rate.

Step 2 : Check the increase, decrease or constant.

2) Detailed Analysis:

In Detailed Analysis, the paper will be showing detailed analysis of the stocks, which will help the user to identify the profitable stocks. It will be including MACD charts, Historic prices chart etc

a)MACD Chart:

MACD(Moving Average Convergence Divergence) is the most popular price oscillator. It compares fast moving average of a series with slow moving average of a series[7]. Fast moving average is a short term moving average. It is more reactive. Short Moving average is a long term moving average. It is more lethargic. It's implementation is as follows:

Step 1: Get historical daily close prices : Getting historical stock quotes from Yahoo Finance.

Step 2: 12-day EMA of the close prices.

The first value is simply a trailing 12-day average, calculated with Excel's =AVERAGE() function. All other values are given by this formula.

$$EMA_n = Closing\ Price_n \frac{2}{Time\ Period + 1} + EMA_{n-1}\left(1 - \frac{2}{Time\ Period + 1}\right) \quad (3)$$

where Time Period is 12, n refers to today, and n-1 refers to yesterday. Essentially, today's EMA is a function of today's closing price and yesterday's EMA.

Step 3: 26-day EMA of the close prices

Again, the first value is simply an average of the last 26 day's closing prices, with all other values given by the above formula (with the Time Period equal to 26)

ISSN:2319-7900

Step 4: Calculate the MACD

The MACD is simply the 12 day EMA minus the 26 day EMA.

Step 5: The signal line

This is a 9-day EMA of the MACD. The first value is simply a 9-day trailing average. All other values are given by this equation, where the time period is 9.

$$signal_n = MACD_n \frac{2}{Time\ Period + 1} + signal_{n-1}\left(1 - \frac{2}{Time\ Period + 1}\right)$$

(4)

b)Historic Prices:

Historic prices of a particular stock shows us stream of its performance in market and daily up's and down's. These can be helpful in knowing performance of a particular stock in past years. Novice and expert users can use this information to gain insight into the stock data. This can be shown using line charts and bar charts.

gvisLineChart() and gvisBarChart() are the built-in functions in googleVis which will be used for visualising the historic prices. Input to these functions will be daily closing prices of a particular stock. Date will be plotted on X-axis v/s Daily closing price on Y-axis.

Proposed Outcome

Initially, the news feeds about a particular stock will be retrieved by the data retriever. These news feeds will be fed to the Message understanding component of the NLP, where the news parsed and will be given to the Named Entity Recogniser . The NER will categorise the tokens and store them in template format. These templates are then given to the Statistical Co-relation Component to test the co-relation of the patterns.

Further, Perceptron Backtracking is done which tells us whether we should buy or sell the stock. And finally, Visualisation of the data is done to predict the stock market.

Conclusion

The stock market is very complex environment to deal with . One of the major facts about the stock market is that it is highly profitable investment but one where there is absolutely no guidance of experienced investors. The stock market is not stagnant. Changes take place rapidly. Our proposed DSS considers all the aspects including the use of Data Analytics of historical prices, the local stock market parameters and the online news feeds , which are often excluded and ignored in many existing systems.

The ultimate aim of our system is to provide some insight to the end user as to why a certain conclusion is reached and help his decision making process. The system assumes that the market is pretty tough to predict and it therefore provides a thorough analysis to the end user to assist him in the final decision.

Merits & Demerits

a) Merits:

1) Our system gives the result on the basis of static data or historical data, Local Stock Parameters and Online News Feeds.

2)Visualisation of Data is done to provide a better insight to the user.

3) Algorithms used are simple, straightforward and are based on the firm foundation of maximizing variances.

b)Demerits:

1) Our system is dependent on R language for analysis.

2) It also depends on several websites for input data.
3)It depends on online tool - KIMONO for data retrieving process.

Acknowledgments

We are thankful to IJACT Journal for the support to develop this document.

References

[1] Ajinkya M. Vaidya, Nikunjkumar H. Waghela, Sneha S. Yewale, “Decision Support System for Stock Market using Data Analytics and Artificial Intelligence, International Journal of Computer Applications(0975-8887) Volume 117- No. 8, May 2015.

ISSN:2319-7900

[2] Fredrick S .M Herz "Stock Market Prediction Using Natural Language Processing " US 200/0135445 A1- Jul. 17, 2003

[3] Markus Gesmann, Diego de Castilo , "Using the Google Chart tools with R-CRAN: googleVis-0.5.10 Package Vignette".

[4] R B. Parihar, R V. Argiddi. (2011) ," An Optimized Approach to Analyze Stock market using Data Mining Technique"- Proceedings published by International Journal of Computer Applications (IJCA) International Conference on Emerging Technology Trends (ICETT).

[5] Sprague, R; (1980). "A Framework for the Development of Decision Support Systems." MIS Quarterly. Vol. 4, No. 4, pp.1-25.

[6] Rosenblatt, Frank (1957), The Perceptron--a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

[7] Aseema Dake Kulkarni and Ajit More, "An application of Moving Average Convergence And Divergence(MACD) indicator on selected stocks listed on Bombay Stock Exchange(BSE).

Biography

S.A.MAHAJAN (Surendra Mahajan) received the Bachelor's degree in Computer Science and Engineering from the University of Amravati, Maharashtra, India, Master degree in Information Technology from Bharati Vidyapeeth University Pune, Maharashtra,India. He is currently an Assistant Professor in the Department of Computer Engineering & Information Technology at P.V.G.'s College of Engineering & Technology, Pune, Maharashtra,India. His research interests are in Software Engineering, Software Testing, Database Management system. Prof.S.A.Mahajan can be reached at sa_mahajan@yahoo.com

ATHARVA .S. DESHPANDE going to receive his Bachelor of Engineering Degree in Information Technology from PVG's College of Engineering and Technology, Savitribai Phule University, Pune.

MANASI .R. AMINBHAVI going to receive her Bachelor of Engineering Degree in Information Technology from PVG's College of Engineering and Technology, Savitribai Phule University, Pune.

ASLESHA .P. BARKE going to receive her Bachelor of Engineering Degree in Information Technology from PVG's College of Engineering and Technology, Savitribai Phule University, Pune.

RAHUL.. BHAGWAT going to receive his Bachelor of Engineering Degree in Information Technology from PVG's College of Engineering and Technology, Savitribai Phule University, Pune.