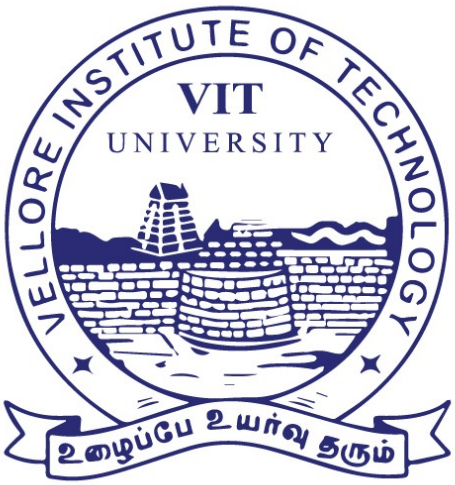


VIT®

UNIVERSITY

(Estd. u/s 3 of UGC Act 1956)

Title : Common man's portfolio management system

Course: Software Design and Developement
CSE 1005

Slot : D1

By
Sheril S. Philip (15BCB0120)
Debashis Karmakar (15BCB0049)
Sammyak Rokade(15BCB0107)

Introduction

1.1 General Instruction:

Out of all the books offering investing advice to research papers analyzing mathematical prediction models, the stock market has always been centre of attraction for public and academic interest. Number of publications propose strategies with good profits, while others demonstrate the random and unpredictable behaviour of share prices. This debate on how to predict stock market recently piqued our interest and led us to choose our Major Project topic within this area of research. The following observations influenced our decision:

- There is large amount of relevant financial data available on the internet which is increasing day by day.
- Large number of C. Sc disciplines including software engineering, databases, distributed systems and machine learning have increased possibility to apply skills.
- The opportunity to expand our knowledge in finance and investing, as we had only little prior exposure to these fields.

The following sections define the goal of the project and give an overview of the system that was built.

Basics:

In order to clarify the goal of the project, following are the dominant schools of thought on investing must first be introduced.

Fundamental analysis

This approach is to analyse fundamental attributes in order to identify promising companies.

This includes characteristics such as financial results, company's assets, liabilities, and stock and growth forecasts. It's very important to understand that this type of analysis is not static; newly released financial information, corporate announcements and other news can influence the fundamental outlook of a company. Fundamental analysis requires expertise in a particular sector and is often conducted by professional analysts. Their recommended investments are regularly published and updated.

Technical analysis

In contrast to fundamental analysis, technical analysis does not try to gain deep insight into a company's business. It assumes the available public information does not offer a competitive trading advantage. Instead, it focuses on studying a company's historical share price and on identifying patterns in the chart. The intention is to recognize trends in advance and to capitalize on them.

Goal

The goal was to build a system capable of the following tasks:

1. **Collecting fundamental and technical data from the internet**

The system should be able to crawl specific websites to extract fundamental data like news articles and analyst recommendations. Furthermore, it should be able to collect technical data in the form of historical share prices.

2. Simulating trading strategies

The system should offer ways to specify and simulate fundamental and technical trading strategies. Additionally, combining the two approaches should be possible.

3. **Evaluating and visualizing trading strategies**

The system should evaluate and visualize the financial performance of the simulated strategies. This allows a comparison to be made between technical, fundamental and the combined approaches.

Financial information sources on the Web.

1. www.yahoo.finance.com
2. Moneycontrol.com- maintain excellent electronic versions of their daily issues.
3. Reuters (www.investools.com)
4. www.nseindia.com

This rich variety of on-line information and news make it an attractive resource from which to mine knowledge. Data mining and analysis of such financial information can aid stock market predictions.

1.2 Relevant current/open problems.

- Data-are-humongous, nowadays we are seeing a rapid-explosion of numerical-stockquotes and textual-data. They are provided from all different-sources.
- Demand forecasts are important since the basic op management process, going from the vendor raw-materials to finished goods in the customers' hands, takes some time. Most firms cannot-wait for demand to elevate and then give a reaction. Instead, they make-up their mind and plan according to future demand so-that they can react spontaneously to customer's order as they arrive.
- Generally, demand forecasts-lead to good-ops-and great-levels of customer satisfaction, while bad forecast will definitely-lead to costly ops and worst-levels of customer satisfaction.
- A confusion for the forecast is the horizon, which is, how distant in the future will the forecast project? As a simple rule, the away into the future we see, the more blurry our vision will become -- distant forecasts will be inaccurate that short-range forecasts.

1.3 Problem statement

As we discussed problems above we are going to implement the following:

- In this project, we-are trying to review the possibility to apply two-known techniques which is neural-network and data-mining in stock market prediction. Extract useful information from a huge amount of data set and data mining is also able to predict future trends and behaviors through neural network. Therefore, combining both these methods could make the prediction much suitable and reliable.
- The most important for predicting stock market prices are neural networks because they are able to learn nonlinear-mappings between inputs and outputs.
- It may be possible to perform better than traditional analysis and other computerbased methods with the neural-networks ability to learn-nonlinear, chaotic-systems.

1.4 Overview of proposed solution approach

- Basically the main objective of this project is to collect the stock information for some previous years and then accordingly predict the results for the predicting what would happen next. So for we are going to use of two well-known techniques neural network and data mining for stock market prediction. Extract useful information from a huge amount of data set and data mining is also able to predict future trends and behaviors through neural network. Therefore, combining both these techniques could make the prediction more suitable and much more reliable.
- As far as the solutions for the above problems, the answer depends on which way the forecast is used for. So the procedures that we will be using have proven to be very applicable to the task of forecasting product demand in a logistics system. Many techniques, which can prove useful for forecasting-problems, have shown to be inadequate to the task of demand forecasting in logistics systems.

Novelty/Benefits:

The rich variety of on-line information and news make it an attractive resource from which one can get data. Stock market predictions can be aided by data mining and analysis of such financial information.
Numerical stock quotes collected from yahoo finance and reuters.com are available in organised manner but we have to apply some techniques to parse textual news information about **Indian stock market** is collected from websites released daily.

1.5 Comparative Study of Prediction Techniques Table 9.

Criteria	Technical Analysis	Fundamental Analysis	Traditional Time Series Analysis	Machine Learning Techniques
Data Used	Price, volume,highest, lowest prices	Growth, dividend payment, sales level, interest rates, tax rates etc.	Historical data	Set of sample data
Learning methods	Extraction of trading rules from charts	Simple trading rules extraction	Regression analysis on attributes is used	Inductive learning is used
Type of Tools	Charts are used	Trading rules	Simple Regression and Multivariate analysis used for time series.	Nearest neighbor and Neural Networks are used
Implementation	Daily basis prediction	Long –term basis prediction	Long –term basis prediction	Daily basis prediction

1.6 Details of Empirical Study:

Collection of Stock quotes, Analyst Advice and News:

The information of stock-market is collected once-a-day for the companies in NSE-&-BSE with-a-database of 1 year.

1. NSE-&-BSE Stock-Index-Dataset: The released data on financial-websites is-divided into two-part numerical-quotes & textual-format news-data.

2. It's collected from yahoo-finance; they are available-on-sites at all the time as in .csv file.

3. Historical-prices of the quotes of stocks-and the daily-published-news are collected.

4. Recommendations of-the-analysts are collected-and analysed collectively.

Numerical Representation

Quotes of stocks-are-normalized by scaling-down its unit so-they-occur in small-ranges of 0.0. to 1.0 .These values-that-are-normalized are input values for the-attributes in the tuples of-training so that to fasten the learning-phase.

News Documents Pre-processing

Pre-processing has a unit which is used-to-process the non-structured news-documents. A priori-domain-knowledge is fed up such as .txt-files of stop-words such as an, the, a, of etc. The module of prediction compromises greatly trained many-layered feed-forward-neural network. Back-propagation is an algorithm-of-learning for neural-network. Speaking roughly, a neural-network is the connected-input-and-output sets of unit in which each-conn has weight-attached with it. During the learning phase, the network-learns by changing-the weights so that correct-class-label of the input-tuples can be predicted

2. Integrated Summary

The most interesting task is to predict the market. So many methods are used for completing this task. Methods, vary from very informal ways to many formal ways a lot. These tech. are categorized as Prediction Methods, Traditional Time Series, Tech Analysis Methods, Mach Learning Methods and Fundamental Analysis Methods. The criteria to this category is the kind of tool and the kind of data that these methods are consuming in order to predict the market. What is mutual to the technique is that they are predicting and hence helping from the market's future behaviour.

Technical Analysis Methods:

Method of guessing the correct time to vend or purchase a stock pricing. The reason behind tech analysis is that share prices move in developments uttered by the repetitively altering qualities of investors in answer to different forces. The tech data such as price, volume, peak and bottom prices per trade-off period is used for graphic representation to forecast future stock activities.

Fundamental Analysis Techniques:

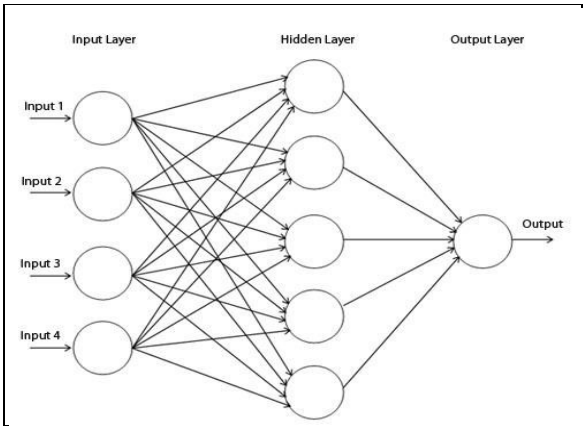
This practice uses the theory of firm foundation for preferred-stock selection. Data of fundamental analysis can be used by forecasters for using this tech of prediction for having a fully clear idea about the market or for investment. The growth, the bonus pay out, the IR, the risk of investing so on are the standards that will be used to get the real value for an asset in which they could finance in the market. Main target of this process is to determine an inherent value of an strength.

Traditional Time Series Prediction: Past data is used here and it uses this data to find coming values for the time series as a linear grouping. Use of Regression depictions have been used for forecasting stock market time series. Two rudimentary types of time series are simple and multivariate regressions.

Machine Learning Methods: The main reason is inductive learning. These types of methods use samples of data that is needed for creating an hope for the underling function that had produced all of the other data. Taking out a deduction from different samples which are given to the model is the main aim for this. The Nearest Neighbour and the Neural Networks Practices have been used for forecasting of the market.

Prediction Module

Multi-layered Feed-Forward network



This neural network has one layer of input, concealed layer, and one yield layer.

Input layer: Made up of units; the qualities measured for each drill tuple matches to the input to the network. Inputs are served to this layer instantaneously. The input passes through input layer and weighted & instantaneously served to the next layer i.e. hidden layer.

Hidden Layer: The productions of the input layer are input to this concealed layer. The number of concealed layer is random; in rehearsal only one concealed layer is used. The weighted output of the concealed layer are input to the next or output layer, which actually releases the network forecast for given tuples.

Output Layer: This layer actually discharges the network forecast for given tuples. Multilayer feed forward network are able to model the class forecast as a nonlinear grouping of the input. For given concealed units and enough preparation samples can carefully estimate to any function

ANALYSIS, DESIGN & MODELING

3.1 Overall description of the project

Project is overall based upon the myraid data which is going to be mined from various stock related portals and after fetching the desired data they have been used for the predictions of related results.

Collection of Stock quotes

We are collecting the stock information once in day.

1. BSE Stock Index Dataset: The data released over financial websites is in both the forms like numerical quotes and textual format news data.
2. It is collected from yahoo/finance; they are available on sites at all the time as in .csv file.
3. Historical prices of the stock quotes are collected.
4. Stock quotes are normalized by scaling its units so they occur in small range of 0.0. to 1.0 . To speed up the erudition phase these normalized values are used as put in values for each feature in the training tuples.

News Documents Pre-processing:

To process unstructured news documents the Pre-processor unit is used. It is fed up with a priori field knowledge such as .txt files of stop words such as a, an, the, of etc.

The following steps involved in pre-processor unit are as follows:

I. Stop Words Removal
II. Stemming
III. Key Phrases Extraction.

These key phrases are initiated with some weight .Lastly, The system assessment on the stocks from India's Bombay Stock Exchange is carried out. For given day's open index, day's high, day's low, volume and adjacent values along with the stock news textual data, our forecaster will forecast the final index price for given trading date.

3.2 FUNCTIONAL REQUIREMENTS:

The prediction shall abide by the following functional requirements:

1. Prior to application of stock recommendations, the database is updated by the latest values.
2. The charts and comparison of the companies would be done only on the latest data stock market data.
3. The user is provided with a login, logging into which enables the user to view his past stock purchases and future recommendations.
4. The user can look previous data Information which was collected.
5. Each user has a friend list and can also be recommended on their buying patterns.
6. The user can also be recommended on the basis of the trending stocks which would require the data regarding the stocks.

3.3 NON FUNCTIONAL REQUIREMENTS:

1. Reliability: The reliability of the product will be dependent on the accuracy of the data- date of purchase, how much stock was purchased, high and low value range as well as opening and closing figures. Also the stock data used in the training would determine the reliability of the software.

2. Security: The user will only be able to access the website using his login details and will not be able to access the computations happening at the back end.

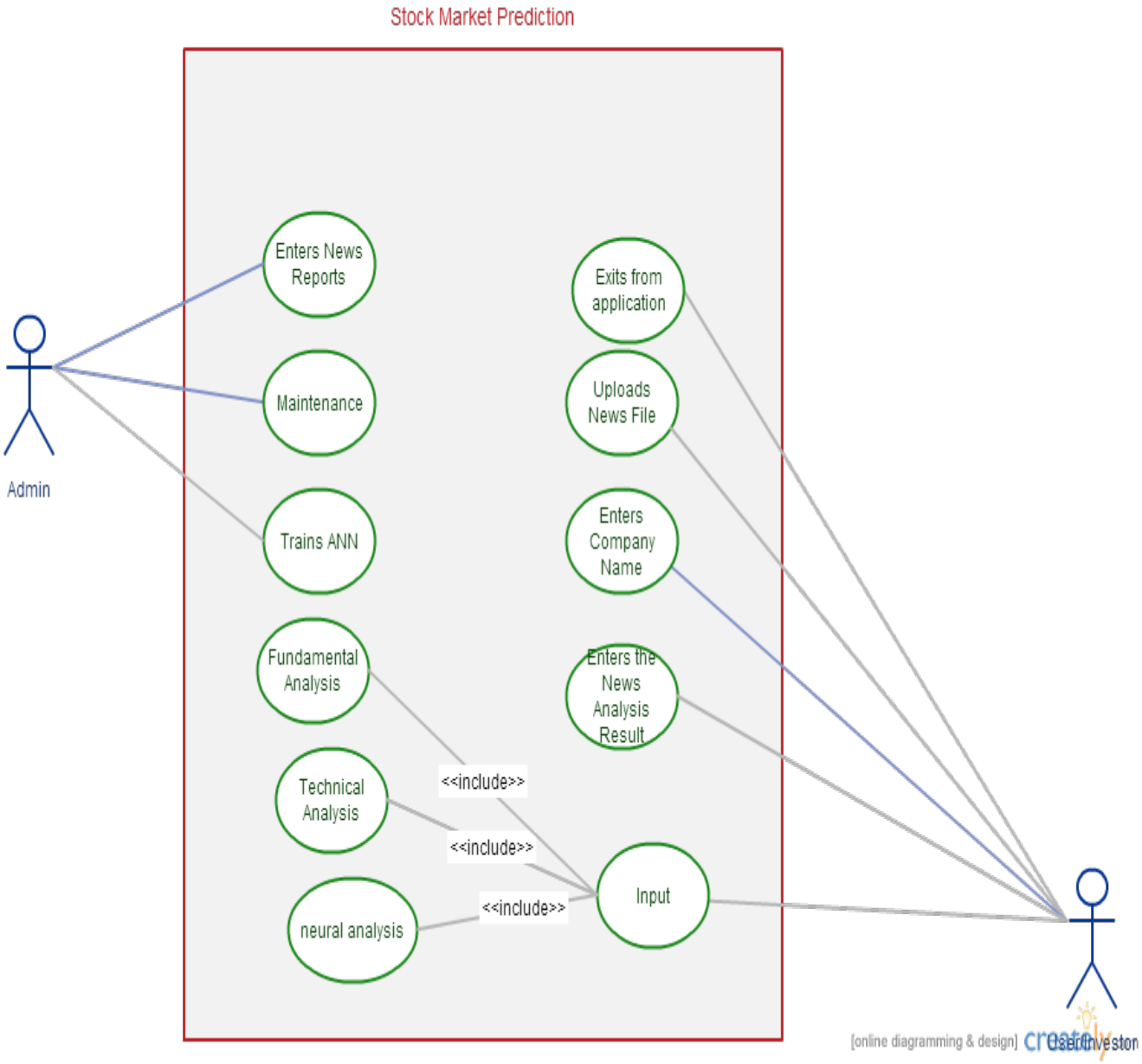
3. Maintainability: The maintenance of the product would require training of the software by recent data so that there commendations are up to date. The database has to be updated with recent values.

4. Portability: The website is completely portable and the recommendations completely trustworthy as the data is dynamically updated.

5. Interoperability: The interoperability of the website is very high because it synchronize all the database with the wamp server.

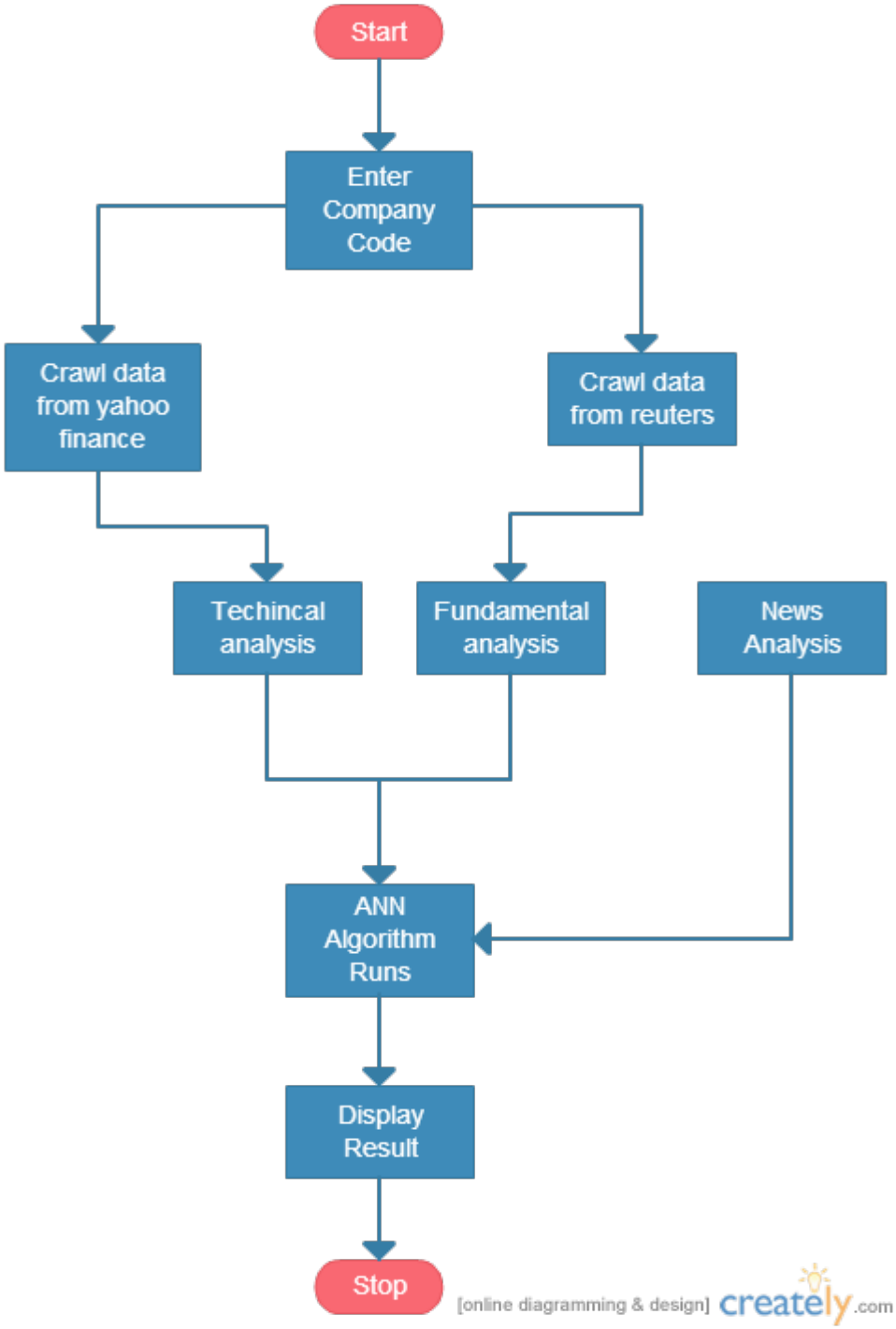
3.4 Design Diagrams

USE CASE DIAGRAM

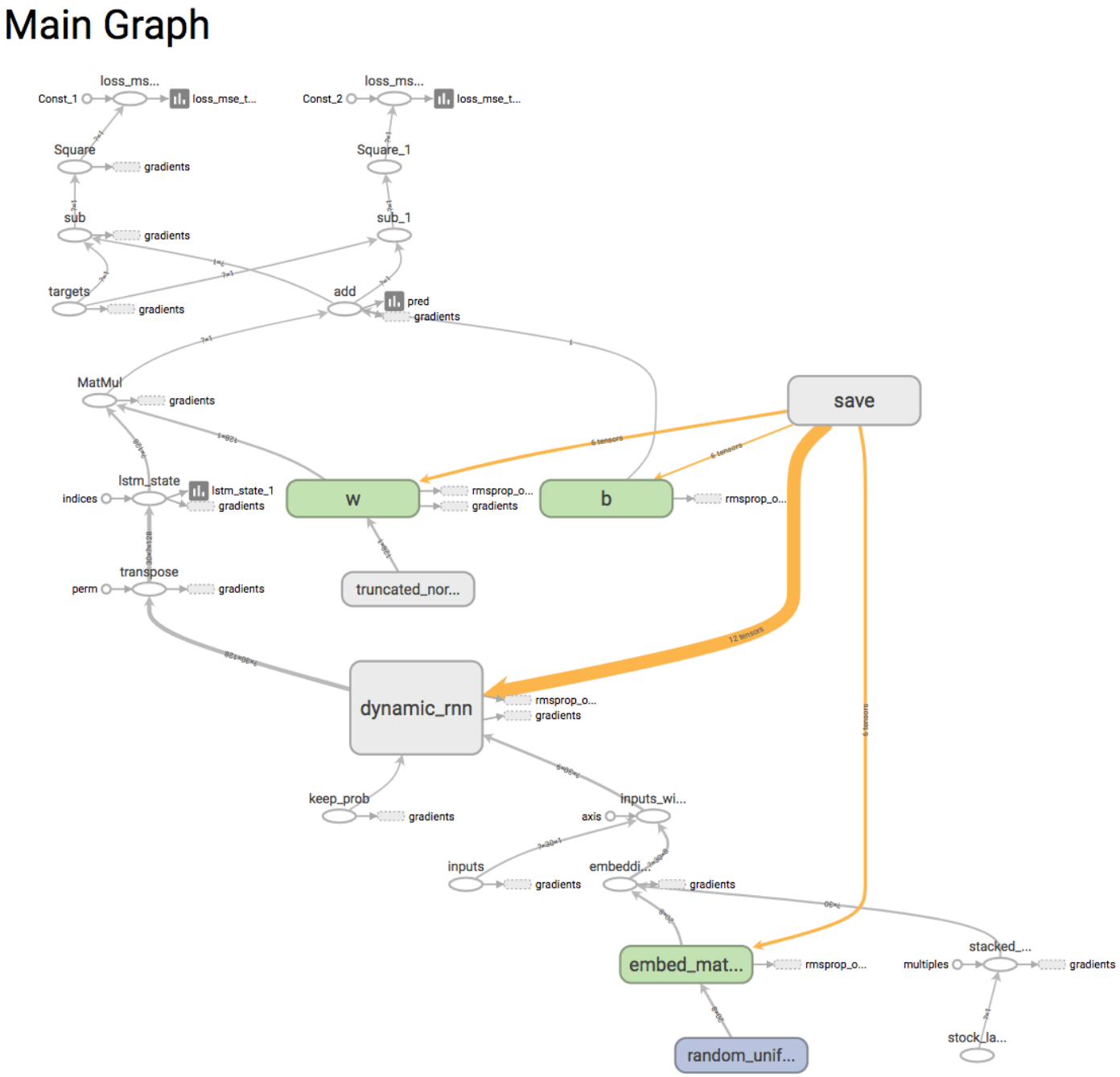


CONTROL FLOW DIAGRAM

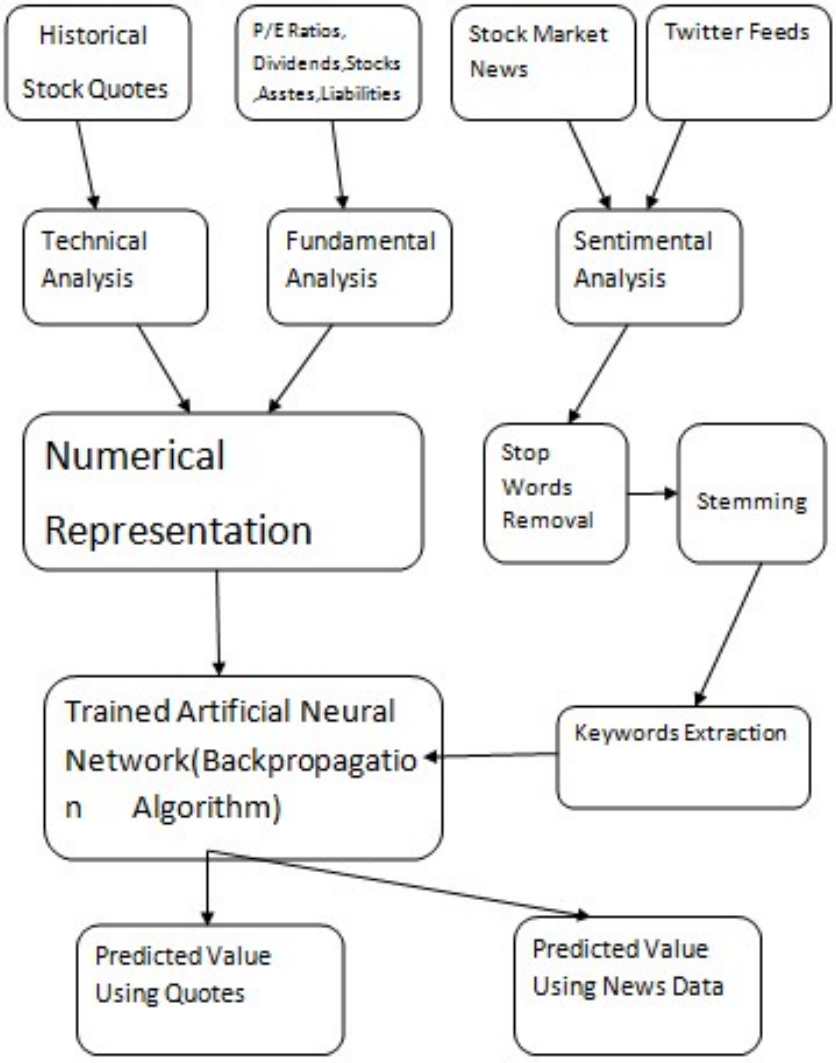
Overall Control Flow Diagram

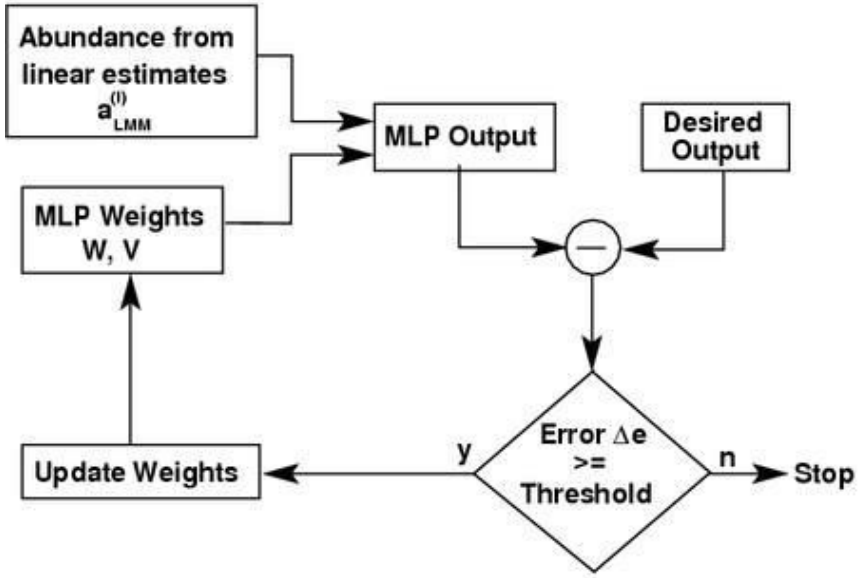


TensorFlow DFD



ACTIVITY DIAGRAM





IMPLEMENTATION AND TESTING

Implementation Details and Issues:

The Crawler - Data Sources

In the early phase, a large no. of websites was considered and the ones most appropriate for the project were identified. The following sections outline characteristics of each data source and list some examples. Data Source - Type URL

Money control → News Fundamental moneycontrol.com

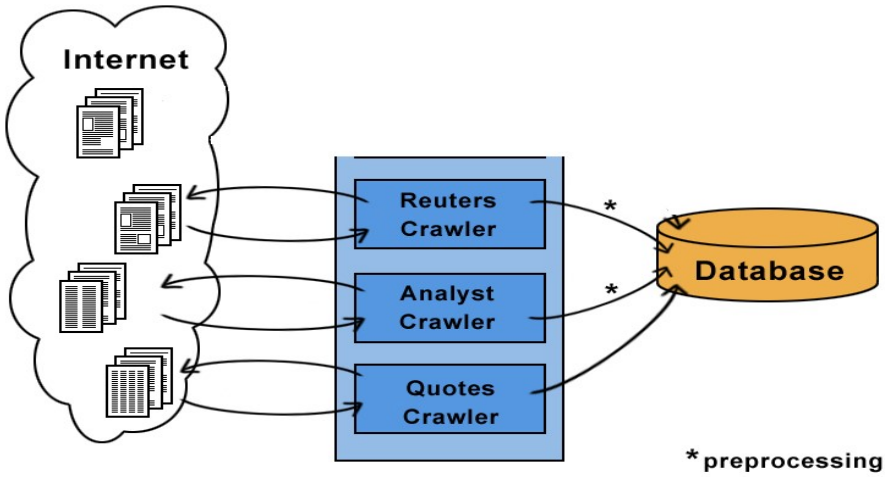
Reuters '→ Reuters Finance Analyst Recommendations

Yahoo Finance → Historical Prices Technical finance.yahoo.com

Yahoo's financial portal includes current and past analyst recommendations for each company. This makes it possible to track the changing sentiment of analysts by following the upgrades and downgrades over time.

Yahoo Finance Historical Prices

After analysing Yahoo Finance, Yahoo's historical stock quotes were selected. The quotes which are selected consists of daily opening, low, high and closing prices and are adjusted for stock splits and dividends. The other sources were also more desirable, but was abandoned because of periods of missing prices and some price inconsistencies when compared to services like Yahoo and Google.



The quotes crawler does not need this phase, as Yahoo's historical quotes are conveniently available in CSV format.

Parsing analyst recommendations

Different research firms tend to use different vocabulary for recommendations. For example, some use Market Outperform, while others use Over-weight or simply Buy to suggest a buying opportunity. In order to compare recommendations, all 96 different phrases found in the dataset were manually mapped to the three expressions Buy, Neutral and Sell. Appendix B lists the various phrases and their mappings.

Computing Trading Signals

The fundamental and technical signals for the evaluating companies are described. A company is deemed potential when all the specified signals point out a rising price trend. Accordingly, a company is deemed failing when all specified signals predict a downward price trend.

Typically, several companies meet the criteria on a given day.

Analyst Recommendations

Due to the pre-processing, the analyst recommendations were easily comparable across research firms. Thus, they could be aggregated to an analyst sentiment. At any given time, the number of analysts recommending Buy, Neutral or Sell could be computed (nBuy, nNeutral and nSell accordingly). This resulted in the following signal:

$$signal = \begin{cases} 1.0 & sentiment > threshold_1, n \geq min \\ 0.0 & sentiment < threshold_2, n \geq min \\ 0.5 & \text{else} \end{cases}$$

$$\text{where } n = n_{Buy} + n_{Neutral} + n_{Sell},$$

$$sentiment = \frac{n_{Buy}}{n},$$

The values threshold1 and threshold2 represent levels of analyst sentiment that must be met to trigger buy or sell signals; e.g. selecting a value of 0.7 for threshold1 means 70% of the analysts must be recommending a Buy. The parameter min specifies the least number of analysts required to compute a signal.

Technical Trading Signals

The technical analysis in detail is covered by the book 'New Trading Systems and Methods'. After studying the book, 2 out of four technical signals seemed promising and were implemented: 1. Moving Average

2. Moving Average Convergence Divergence
3. Relative Strength Index
4. Stochastic

To give the reader a favour of technical analysis, the Moving Average and RSI will be explained in the subsequent sections.

Moving Average

A moving average is a effortless technique to recommend buying and selling points on a stock price chart. For this purpose, the mean share price in a trailing window is calculated. Common values for the window size are 20 days, 60 days and 200 days. When the present prices rise beyond the moving

average, a procure signal is triggered. A sell signal is triggered when the present price comes down below the moving average.

With pt symbolizing the share price at time t, a moving average signal can simply be expressed as:

$$signal = \begin{cases} 1.0 & p_t > movingAverage(n) \\ 0.0 & p_t < movingAverage(n) \\ 0.5 & \text{else} \end{cases}$$

$$\text{where } movingAverage(n) = \frac{1}{n}\sum_{i=1}^{n} p_{t-i}$$

Combining Trading Signals

A trading strategy can use one or more of the signals. When using more than one signal, a scheme for combining them is required. The following sections describe the two possible combination techniques that were implemented.

Simple Combinations

A simple way to combine the output of several signals is to only signal a buy or sell when all specified signals do so. This can be expressed as follows:

$$\begin{cases} 1.0 \text{ if all individual signals return 1.0} \\ \text{signal (combined)} = \\ 0.0 \text{ if all individual signals return 0.0} \\ 0.5 \text{ else} \end{cases}$$

Combinations using Neural Networks

Using historical data, a neural network can be learned that describes how trading signals are related to subsequent price movements. Figure shows how the implemented fundamental and technical signals can be used as input values and how the expected future price trend is the desired output. The trained neural network can then be used on new data to forecast future price movements and make investments. Details are explained in the following sections.

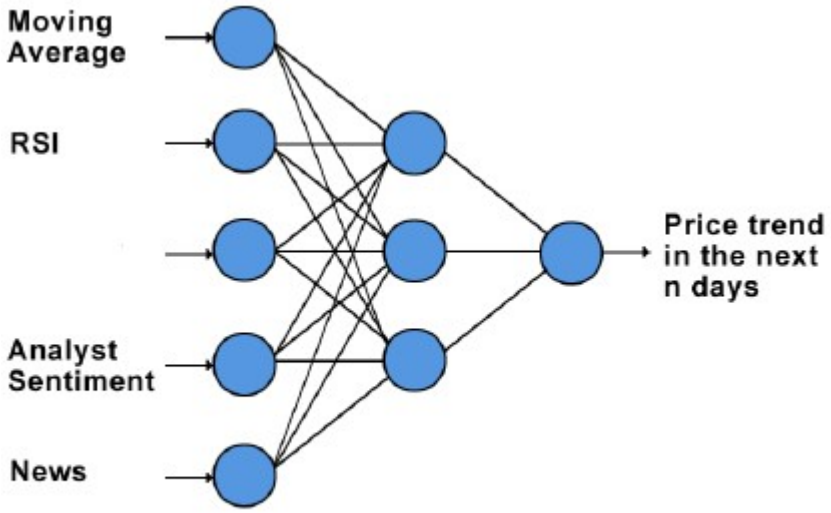


Figure 3.4: The neural network setup

The input values are all normalized to the continuous [0,1] range.

Output

The output is the expected price change in a window of days. The value is in the continuous [0,1] range with 1.0 representing a 10% rise, 0.5 representing no price change and 0.0 representing a 10% price drop.

Training

The neural network's weights are learned using the backpropagation algorithm with a configurable learning rate and number of epochs.

Trading

Once a neural network is built, it can be used by inserting current technical and fundamental input values and computing the predicted output value. If the output crosses a certain upper threshold (e.g.: 0.7), an upward price trend can be predicted and shares can be bought. Likewise, a descending price tendency can be signalized by an yield value below a lower cut-off value and short-selling can take place.

Proposed Algorithm

The study used three-layer (one hidden layer) multilayer feed-forward neural network model taught with backpropagation algorithm.

Backpropagation, an abbreviation for "backward propagation of errors", is a common technique of teaching artificial neural networks used in combination with an optimization method such as gradient descent. The method calculates the gradient of a loss function with respects to all the weights in the network. The gradient is fed to the optimization method which in turn uses it to apprise the weights, in an attempt to minimalize the cost function

The backpropagation learning algorithm can be divided into two phases: propagation and weight update.

Phase 1: Propagation:

"Each propagation involves the following steps:

1. Onward propagation of a training pattern's input through the neural network in order to create the propagation's output initiations.
2. Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the summation of all output and concealed neurons."

Phase 2: Weight update:

For each weight-synapse follow the following steps:

1. Multiply its output summation and input activation to get the gradient of the weight.
2. Subtract a ratio (percentage) of the gradient from the weight.

This proportion (percentage) influences the speediness and worth of learning; it is called the *learning rate*. The bigger the ratio, the quicker the neuron trains; the lesser the ratio, the more precise the training is. The sign of the gradient of a weight specifies where the fault is increasing, this is why the weight must be updated in the opposite direction.

Repeat phase 1 and 2 until the performance of the network is satisfactory.

Risk Analysis and Mitigation Plan

1. Since, we are making software which involves updations/modifications, heavy computations and to and-fro activity may be required. But, the software should never take more than reasonable amount of time, which is the goal of the project. Although, it's a risk if it takes much time.

Probability: Low (1)

Impact: High (5)

2. There are light and background constraints for the application. Due to unavailability of resources or server, we might not be able to use the application.

Probability: High (5)

Impact: High (5)

3. We will never be able to check if our code is upto the requirement, for complex objects, as we will have to update/ modify time to time. Although, this risk can be reduced by machine learning by implementing automatic database updating of new scanned objects.

Probability: Low (1)

Impact: High (5)

Algorithm for a 3-layer network (only one hidden layer):

initialize network weights (often small random values)
do forEach training example ex
prediction = neural-net-output (network, ex) *// forward pass*
actual = teacher-output(ex)
compute error (prediction - actual) at the output units
compute Δw_h for all weights from hidden layer to output layer *// backward pass*
compute Δw_i for all weights from input layer to hidden layer *// backward pass continued*
update network weights *// input layer not modified by error estimate*
until all examples classified correctly or another stopping criterion satisfied
return the network

Testing and requirement analysis

SOFTWARE REQUIREMNTS:

- Python 3.6
- Tensorflow 0.1.4
- Pandas
- Any normal browser to deploy TensorBoard

HARDWARE REQUIREMENTS(Minimum):

- 2.7 GHz Intel Core i5 -- Processor
- 8 GB 1867 MHz DDR3 -- Memory

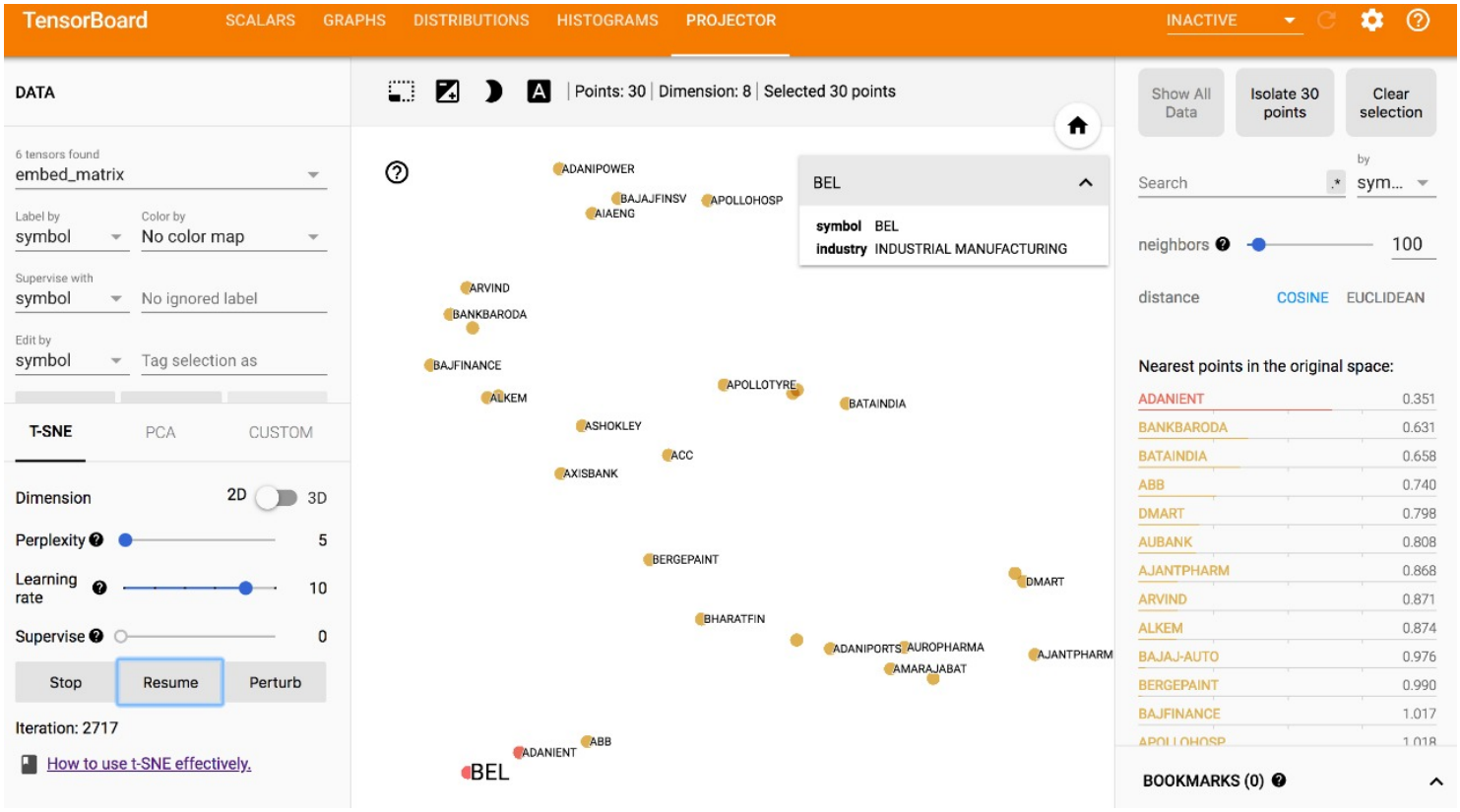
Component decomposition and type of testing required

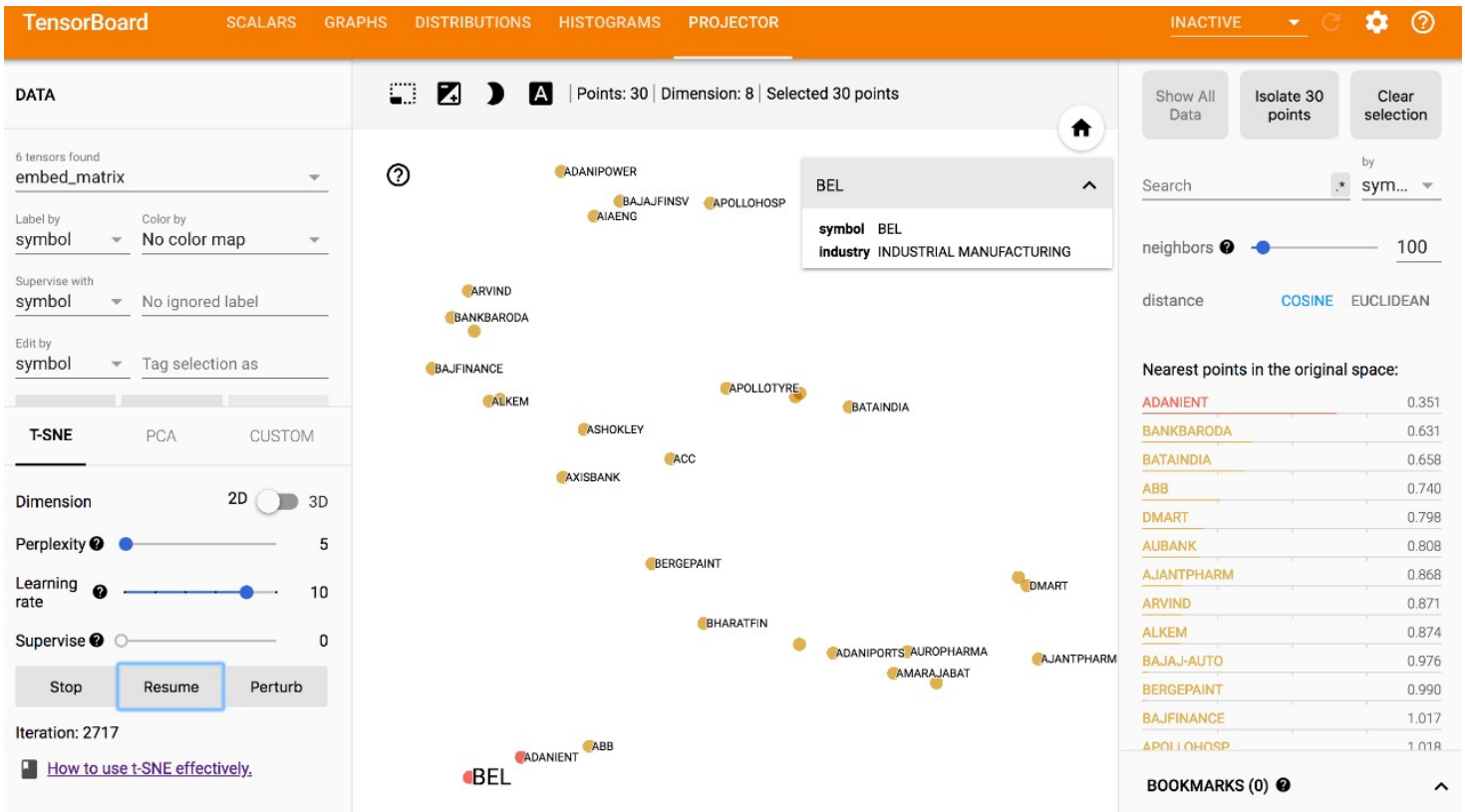
S No.	List of Various Components (modules) that require testing	Type of Testing Required	Technique for writing test cases
1	Data (Stock Price)	Requirement	Black Box(Boundary Values)
2	Algorithms (BP)	Unit	White Box
3	Neural Network (output	Unit	White Box
4	Graphs and Table	Volumes	White Box

Screenshot :

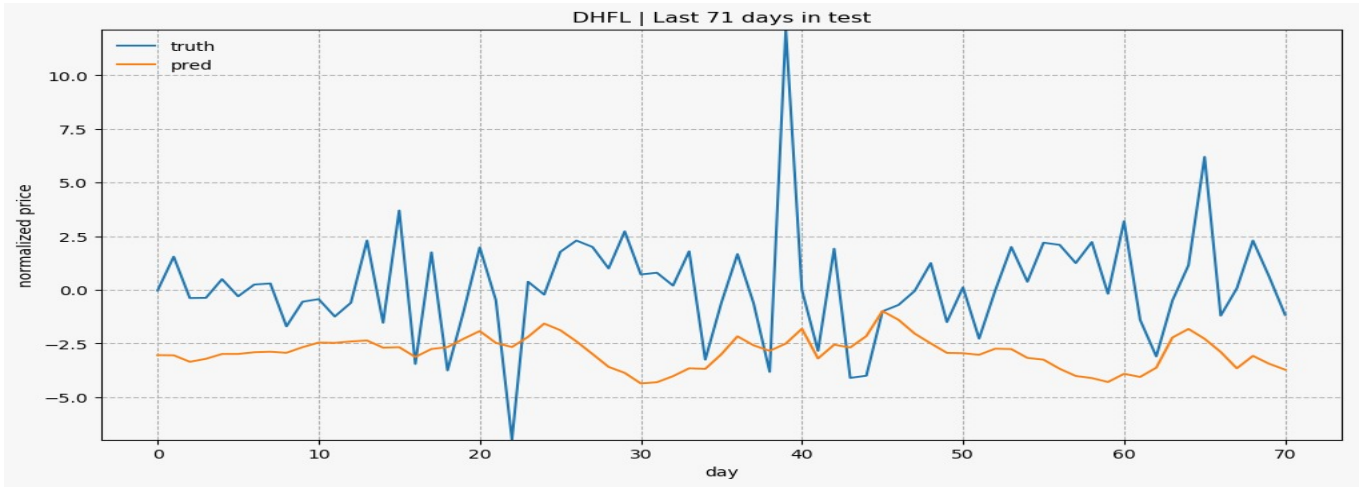
Training :

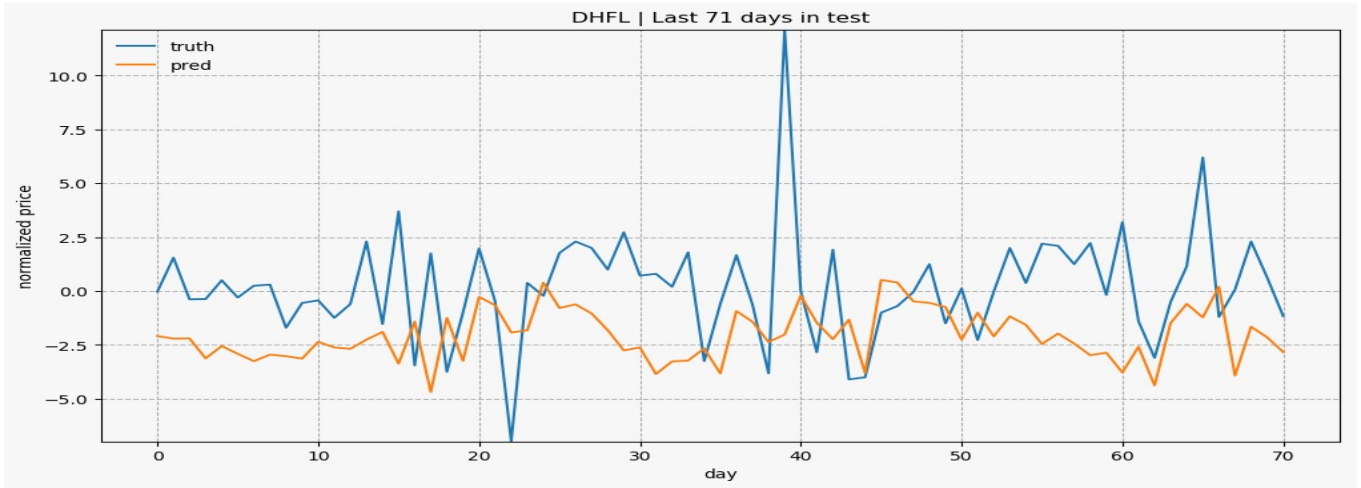
Tensorboard display of stochastic distribution :

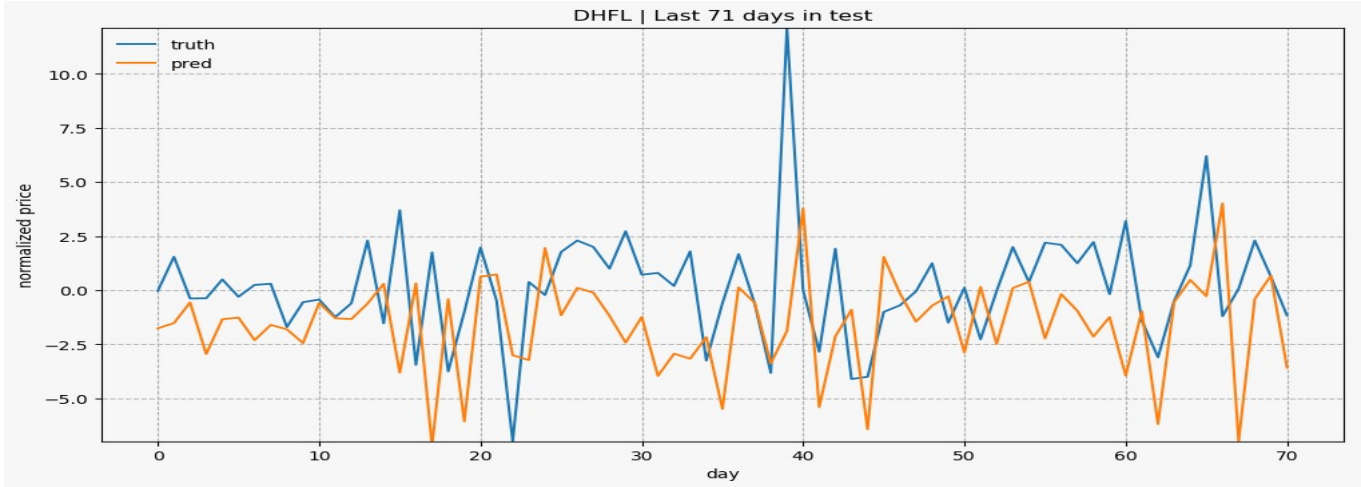


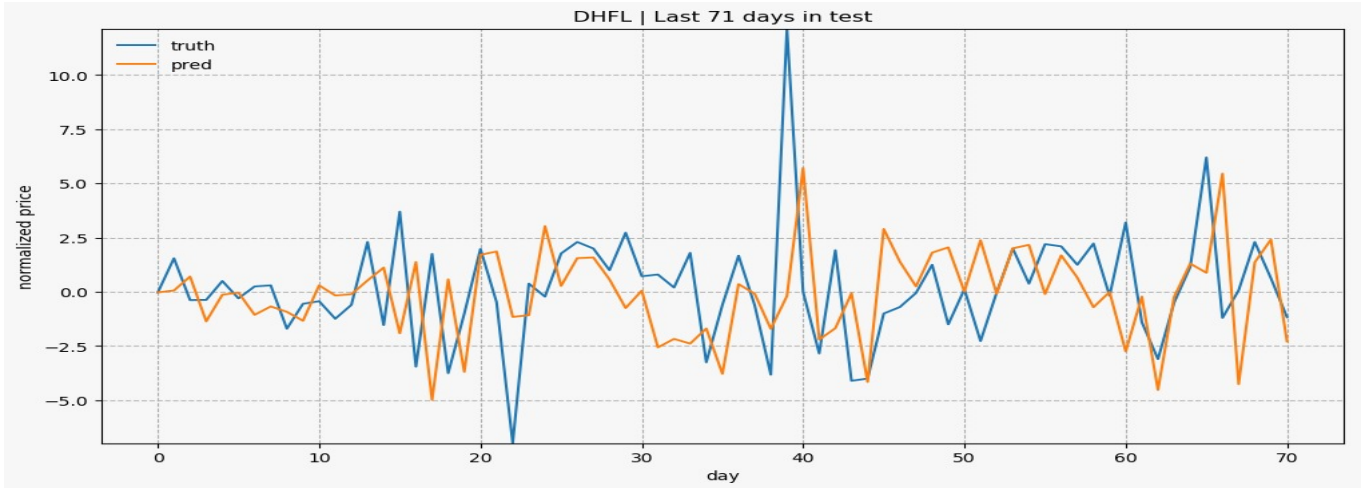


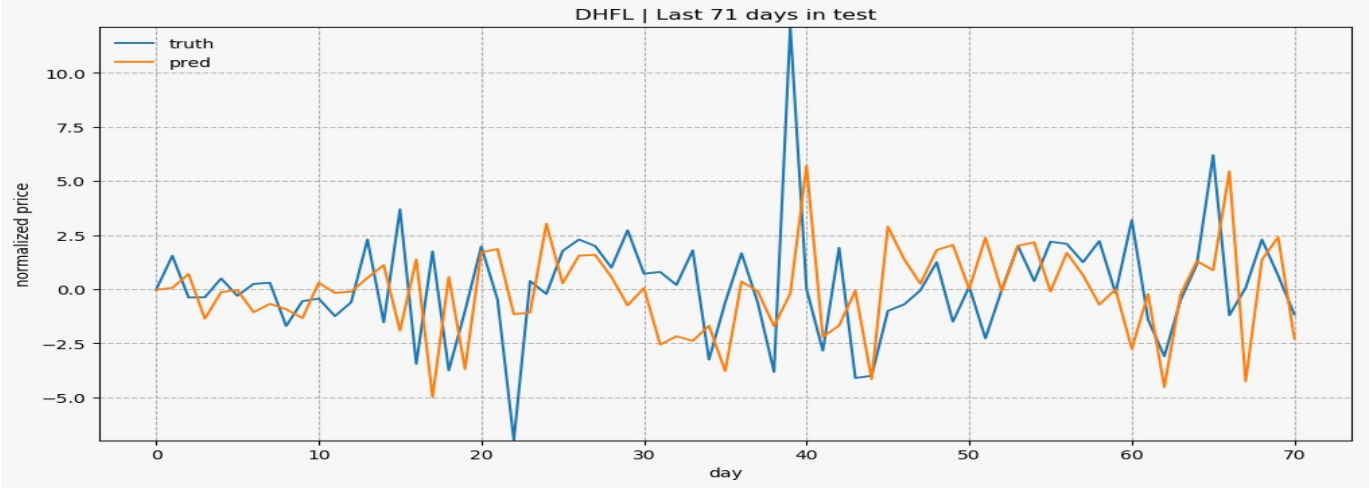
Given Below is the stock name DHFL's Prediction on over 50 epochs :


DHFL | Last 71 days in test
truth
pred
normalized price
day


DHFL | Last 71 days in test
truth
pred
normalized price
day


DHFL | Last 71 days in test
truth
pred
normalized price
day


DHFL | Last 71 days in test
truth
pred
normalized price
day



Limitations of the Solutions

The solution has a few limitations which are relevant to the proper functioning of our application:

1.) Parsing research firms :Several notations were being used for the same research firm (e.g. CSFB and CS First Boston). A map was manually created to ensure the different expressions were mapped to the same firm.

2.) Parsing analyst recommendations: Different research firms tend to use different vocabulary for recommendations. For example, some use use Over-weight or simply buy to suggest a buying opportunity. In order to compare recommendations, all 96 different phrases found in the dataset were manually mapped to the three expressions Buy, Neutral and Sell.

3.) When working on a large project, small bugs can creep in and easily go unnoticed for some time (e.g. array indices of by one). Particularly when running simulations, the results may be greatly affected and the error may be hard to track down. In order to prevent this to a certain extent, unit tests were written using the JUnit4 framework. The behaviour of all relevant simulation server classes could be checked; when refactoring parts of the server, the behaviour could be revalidated.

CONCLUSION

Evaluating the Stock market prediction has at all times been tough work for analysts. Thus, we attempt to make use of vast written data to forecast the stock market indices. If we join both techniques of textual mining and numeric time series analysis the accuracy in predictions can be achieved. Recurrent neural network is qualified to forecast BSE market upcoming trends. Financial analysts, investors can use this prediction model to take trading decision by observing market behaviour.

FUTURE WORK

- More customized model to adapt to stock data.
- Twitter feeds message board, Extracting RSS feeds and news, for sentiment analysis.
- Considering internal factors of the company likes Sales, Assets etc.

References

[1] Daily Stock Market Forecast from Textual Web Data
W• uthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J.; Lam, W.
IEEE International Conference on Systems, Man, and Cybernetics, vol.3, pp.2720-2725 vol.3, 11-14 Oct 1998

[2] The Apache httpclient library is an open source Java library for working with HTTP. *http://hc.apache.org/httpcomponents-client/*

[3] NekoHTML is an open source Java library for fixing HTML. *http://nekohtml.sourceforge.net*

[4] The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data Ronan Feldman and James Sanger Cambridge University Press, 11 Dec 2006

[5] New Trading Systems And Methods Perry J. Kaufman Wiley, 4th Edition, 28 Feb 2005

[6] Stock Market Prediction with Backpropagation Networks Freislben, B. Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, vol.604, pp.451-460, 1992

[7] An Intelligent Forecasting System of Stock Price Using Neural Networks
Baba, N.; Kozaki, M. International Joint Conference on Neural Networks, vol.1, pp.371- 377, 1992