Stock Price Prediction using Quotes and Financial News

Manisha V. Pinto, Kavita Asnani

Abstract— This paper provides 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 [1], 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.

Index Terms— Stock Market, Dow Jones Industrial Average, Key Phrase Extraction Algorithm (KEA), Neural Network, Back Propagation Algorithm

I. INTRODUCTION

Data mining is well founded on the theory that the historic data holds the essential memory for predicting the future direction. 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.

The Efficient Market Hypothesis (EMH), as stated by Fama ([14], [15], [16]), assumes that 'Stock prices fully reflect all their relevant information at any given point in time'. Previous researches have also proved that there is a strong correlation between the time at which news articles are published and the time when the stock prices fluctuate ([17]).

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. In this paper, we describe such an application driven data mining system for stock market prediction.

Manuscript Received October 28, 2011.

Manisha V. Pinto, Department of Information Technology (M.E.), Padre Conceicao College of Engineering, Goa University, Verna, India 09823616506 (e-mail: manisha.pinto@gmail.com).

Kavita Asnani, Department of Information Technology (M.E.), Padre Conceicao College of Engineering, Goa University, Verna, India 09326103205 (e-mail: kavitapcce@gmail.com).

For developing such a prediction model, we make use of the daily prices and the time-stamped news articles corresponding to Dow Jones Industrial Average (DJI) index, collected over the period from November 2007 – March 2008 and August 2010 - May 2011. The news articles are pre-processed and then replaced by their corresponding key phrases. The Key Phrase Extraction Algorithm (KEA) [1] is used for extracting the most influential key phrases pertaining to each news article.

The relationship between the news articles and the trends on the stock prices will be used to train the Artificial Neural Network using the Back propagation Algorithm. The trained Neural Network uses the opening index value and the news articles of the day and predicts the closing value and the probable reason for the same before the close of the trading day. The Stock Market under study is the Dow Jones Industrial Average (DJI).

The rest of the paper is organized as follows. In section II we describe the techniques and architecture needed to make the system to be operationally useful. Section III describes the evaluation of the system in a simulated environment. Section IV summarizes the results, opens further research issues and concludes the paper.

II. SYSTEM DESIGN

A. Review System Architecture

The system consists of three main components: a corpus, a simulation server and a client interface. Figure 1 visualizes the system architecture and interactions between the components. Basic descriptions of the individual components are listed below:

1. Corpus

The Corpus is a collection of the news articles collected for each trading day along with the corresponding quotes. This combination of news and quotes will be used by the simulation server for the purpose of prediction of the stock price.

2. Simulation Server

The simulation server loads the information from the Corpus and runs the prediction strategies when instructed by the client. The server's logic is written in Java and is equipped with unit tests.

3. Client

The client provides the users with a browser-based graphical interface to specify strategies to simulate. The results returned by the simulation server are then visualized appropriately.

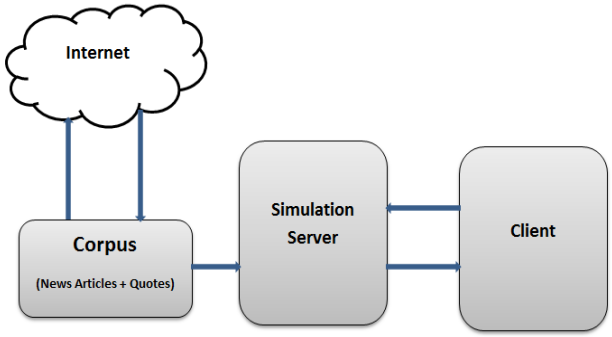

Figure 1: System architecture

B. Block Diagram

The detailed block diagram of the proposed system is given in Figure 2.

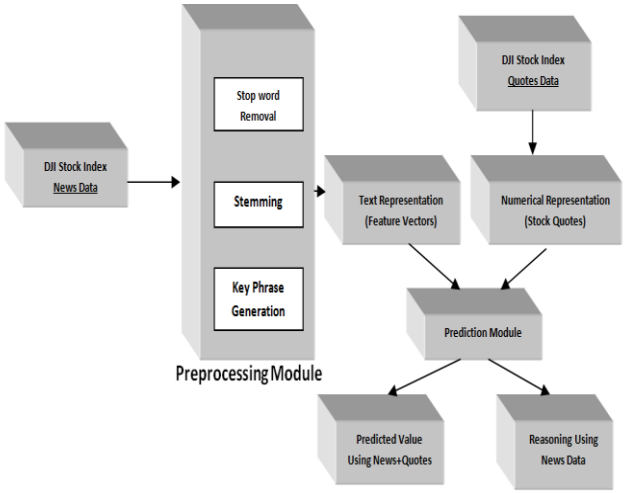

Figure 2 Detailed Block Diagram

The system goes through the following six steps:-

1. DJI Stock Index (News Data) : The data pertaining to DJI index is collected at three consecutive time-intervals during the day,
   - When the stock market begins
   - Mid-day session
   - Before the stock market closes

2. Pre-processing Module: Comprises of the following:
   - Stop Word Removal: The stop word list contains the common words such as conjunctions, articles, particles, prepositions, pronouns, anomalous verbs, adjectives, and adverbs.
   - Stemming: Stemming enables to obtain the grammatical root of a word thus allowing us to treat different variations of a word as the same. E.g. training, trained, train to be all reduced to the grammatical root 'train'.
   - Key phrase Extraction: Generation of the candidate key phrases.

3. DJI Stock Index (Quotes Data): Download the DJI quotes data from www.finance.yahoo.com for the specific time frame.

4. Text Representation (Feature Vectors): News articles are an unstructured form of data with enormous amount of useful information embedded in it. KEA [1], an algorithm for automatically extracting key phrases from text is used.

5. Numerical Representation (Stock Quotes): An attribute is normalised by scaling its values so that they fall within a small specified range, such as 0.0 and 1.0. Min-Max normalisation is used.

6. Prediction Module: This is the most important module which will take in the feature vectors pertaining to the news articles and the stock quotes. It will perform two functions:
   - Predicted Value (News data + Quotes data): Will predict the closing value of the DJI index for that day.
   - Reasoning (News data): Provides a reason for the prediction.

C. Creation of Global List of Key Phrases

After all the news articles contained in the corpus have been pre-processed and the .key files (containing the key phrases) corresponding to each of the .txt files have been generated, a global list of key phrases is created. The attempt here is to obtain the top 20 most significant key phrases impacting the entire corpus. The following algorithm is used for the creation of the global list from which the top 20 key phrases are then obtained.

1. Unite the 5 key phrases from each news article to obtain a single global list. Each key phrase i has a weight, wi,j, in article j, which is an overall probability value provided by the KEA [1] model.

2. Let Wi be the overall weight of keyword i in the global corpus, where Wi = Summation (wi,j).

3. Use Wi to re-rank the list in descending order to select the top 20 key phrases for the corpus.

D. Creation of Global List of Key Phrases

The details of the Neural Network training module are specified in Figure 3.

The network configuration details are as follows:

1. Input layer:

There is only one input layer. In our configuration the input layer has 25 neurons: 5 neurons for the stock quotes data and 20 neurons for the stock text data as follows:

- Neurons 1 to 5 → Normalised stock quotes values: Open, High, Low, Volume and Change (Adj. Close – Close).
- Neurons 6 to 25 →Boolean value i.e. Presence/Absence of the top 20 global key phrases in the key phrases extracted for the news articles corresponding to a given trading day.

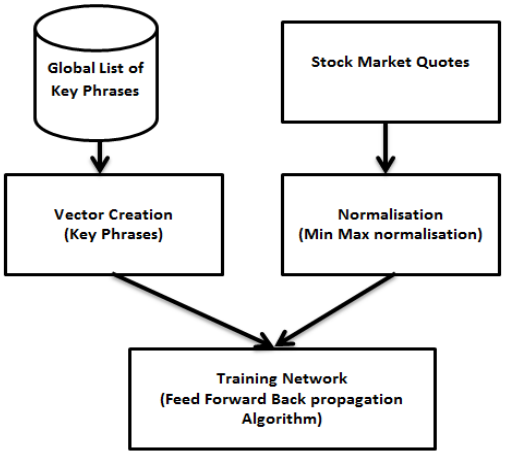

Figure 3 Details of Neural Network Training Module

2. Hidden layer:

The network has only one hidden layer and this layer contains 30 neurons. Each of these neurons will pass on their output to the neurons in the output layer.

3. Output layer:

There is only one output neuron which will contain the Closing index value as predicted by the neural network for a given trading day.

4. Learn Rate

A learn rate of 0.2 has been used.

5. Terminating conditions

Training stops when either

- A pre-specified number of epochs has expired, or
- The error calculated is below a pre-specified threshold.

E. Neural Network Prediction Module

The prediction module comprises of the neural network trained using the back –propagation algorithm. Given a day's open index, day's high, day's low, volume traded and the adjusted close values (in normalised form) along with the stock news data, the predictor module will predict the closing index value for a given trading day. The above specified inputs correspond to the data that is observed after about two hours from the time the stock market opens. The predicted value is then de-normalised to obtain the actual close index value.

F. Reasoning Module

This module is responsible for determining the probable reason behind the stock trend that has been predicted by the neural network. The input data comprises of the vectors that have been generated from the .key files obtained after applying KEA [1] to the news articles collected for each trading day. The vector generation is done using Vector Space Model.

The Vector Space Model calculates the cosine similarity between the news vectors. The similarity is calculated between the news vector corresponding to the prediction day and the previous news vectors. Prior to the similarity calculation, the change in index field is used to obtain a reduced subset of the vectors. For e.g., if the prediction day indicates a positive trend we restrict the application of the cosine similarity only to those trading days which also indicate a positive trend.

After calculating the cosine similarity, we take into consideration only the vector that has a high similarity to the prediction day vector. From the key phrase list of that vector we then obtain the top 4 key phrases. These 4 key phrases will be displayed as the reason for the stock trend.

III. SYSTEM PERFORMANCE

The system was evaluated on stocks from US Dow Jones Industrial Average index on the financial stock data between the time frame of November 2007 – March 2008 and August 2010 - May 2011. The news data was collected from the financial web sites: http://www.finance.yahoo.com, http://reuters.com and www.Marketwatch.com. As mentioned above, the news data was collected at three consecutive periods during each trading day. The stock quotes corresponding to each trading day were downloaded from http://finance.yahoo.com.

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.

From the test data collected over a period of 50+ trading days the following has been concluded:

(i) The Neural Network predictor module provided the correct trend signal for the DJI stock when the difference (Actual Close - Actual Open) lies outside the bound [-82, 50].

(ii) When the difference (Actual Close - Actual Open) lies within the bound [-82, 50], the trend signal for the DJI stock provided by the Neural Network predictor module is not very stable i.e. smaller the gap between the closing price and the opening price, the predictor found it difficult to predict the correct trend and magnitude.

(iii) Out of 51 trading days, the Neural Network Predictor module, correctly determined the trend for 38 trading days, thus giving an efficiency of 74.5%.

(iv) On an average, the price difference between the actual closing index and the predicted closing index for a given trading day was found to be in the range of [-35, 35].

(v) Classifying the magnitude difference (Actual Closing - Predicted Closing) into buckets of 10 units each e.g. [0-10), [10-20), [20-30)....,[90-100), the following was observed:

e.g. 10.3 would be inserted into [10-20).

[0-10) --> 15
[10-20) --> 13
[20-30) --> 14
[30-40) --> 16
[40-50) --> 7
[50-60) --> 8
[60-70) --> 3
[70-80) --> 8
[80-90) --> 4
[90-100) --> 1

Thus, we can see that the magnitude difference (Actual Closing - Predicted Closing), majorly lies in the buckets, [0-10), [10-20), [20-30) & [30-40).

The Global list containing the 20 most influential key phrases from the news articles in the corpus is shown in Table 1.

Table 1 Global list consisting of top 20 key phrases

'opens lower'
'jobless claims'
'open higher'
'slumping dollar'
'battered financial'
'oil futures closing high'
'unemployment rate'
'oil prices'
'lifted financial shares'
'credit crisis'
'strong gains'
'jobless claims data'
'drop in crude oil futures'
'economic news'
'investors\' confidence'
'quarterly loss'
'nuclear power'
'investors\' appetite'
'nuclear crisis'
'required reserves'

IV. CONCLUSION AND FUTURE DIRECTION

This paper is an attempt to determine whether the DJI market news in combination with the historical quotes can efficiently help in the calculation of the DJI closing index for a given trading day. As observed, the DJI closing index value predicted by this application is in close proximity to the actual DJI closing index for a given trading day.

This prediction algorithm will be beneficial to financial analysts who invest in stocks, the investors will be able to foresee the predicted behaviour of their stocks, when relevant news articles are released and take suitable actions depending on the prediction of the stock movement.

The prediction network can be used by any individual, as it will give them a good idea about the trend as well as the unit price of their stocks, thus enabling them to make profits by selling/buying their stocks at the right time.

REFERENCES

1. Ian H. Witten,* Gordon W. Paynter,* Eibe Frank,* Carl Gutwin† and Craig G. Nevill –Manning, KEA: Practical Automatic Keyphrase Extraction
2. Petr Kroha, Thomas Reichel and Bj¨orn Krellner, Text Mining for Indication of Changes in Long-Term Market Trends
3. Robert P. Schumaker, An Analysis of Verbs in Financial News Articles and their Impact on Stock Price
4. Ramon Lawrence, Using Neural Networks to Forecast Stock Market Prices
5. K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam, Financial Stock Market Forecast using Data Mining Techniques
6. Garth Garner, Prediction of Closing Stock Prices
7. Mateusz, KOBOS, Jacek and Mańdziuk, Artificial Intelligence Methods In Stock Index Prediction With The Use Of Newspaper Articles
8. Manoel C. Amorim Neto, Victor M. O. Alves, Gustavo Tavares, Lenildo Arag˜ao Junior, George D. C. Cavalcanti and Tsang Ing Ren
9. Stock Price Forecasting Using Exogenous Time Series and Combined Neural Networks
10. Gil Rachlin' ,Mark Last' , Dima Alberg' and Abraham Kandel2
11. ADMIRAL: A Data Mining Based Financial Trading System,
12. Marc-André Mittermayer, Forecasting Intraday Stock Price Trends with Text Mining Techniques
13. Moshe Koppel and Itai Shtrimberg, Good News or Bad News? Let the Market Decide
14. M. I. Yasef Kaya and M. Elif Karsl_gil, Stock Price Prediction Using Financial News Articles
15. E.F. Fama, Long Term Returns and Behavioral Finance, Social Science Research Network.
16. E.F. Fama, Efficient Capital Markets: A Review of Theory and Empirical Work, Journal of Finance, 25 (May 1970): 383-417.
17. E.F. Fama, Efficient Capital Markets: II, Journal of Finance, 46 (December 1991): 1575-1617.

AUTHORS PROFILE

Manisha V. Pinto received her B.E Degree in Computer Science from Goa University, India in 2004 and her M.E. Degree in Information Technology from Goa University, India in 2011. She has around 4+ years of valued IT Industry experience with extensive work with Java related technologies. Her research interest includes Neural Networks, Stock Market and Finance domain. She is currently based in London, United Kingdom. E-mail: manisha.pinto@gmail.com

Kavita Asnani is Head of Department (Incharge) of Information Technology Department at Padre Conceicao Engineering College, affiliated to Goa University, Goa. She received her Masters degree in Information Technology from Goa University, Goa, India. She has 12 years of teaching experience at College level. She has published many papers in International and National Journals, and also at International and National Conferences. Her area of research includes Data Mining, Information Retrieval and Distributed Systems. E-mail: kavitapcce@gmail.com