Extracting Unstructured Data and Analysis and Prediction of Financial Event Modeling

B.Tharaniya
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka
tharaniyabala@gmail.com

Chethana Liyanapathirana
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka
chethana.l@sliit.lk

Lakmal Rupasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka
lakmal.r@sliit.lk

Kalpa Kalhara Sampath
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka
kalpa.s@sliit.lk

***Abstract*—The stock market is full of uncertainty. It always tends to change with internal and external factors. Since it's commonly accepted that the failure in the Effective market hypothesis, in this research try to focus on the few external factors. Stock price movements are driven by publications of financial news. Investors' or sellers' decisions are made based on the available information considering how a stock market will going up or going down for next few days. News articles incorporate information about a firm's fundamentals, the activities in which a firm is involved and the expectations of other market participants about future price changes. This research tries to focus on the few external factors. Predicting the future expecting variations in the stock market can be by analyzing. This analysis can be done on three types. Those are fundamental, technical and machine learning. In the research paper use the technical analysis. NLP use to Data extraction in the means of a summery, from a selected website. Then clustering and classification into external factor using K-means and KNN. The impact of the extracted summery for the stock market prediction is analyzed using sentiment analysis. Here it will analyses how each external factor implies on the selected sectors in the stock market, therefore it can help in the prediction of actual stock movement for the sectors. Prediction based on Naïve Bayes, Recurrent Neural Network (RNN) approach and the sentiment score of the text segment is calculated by sentiment analysis of the text through SVM.**

Keywords- Naïve Bayes, RNN, K-Means, KNN, NLP, SVM, sentimental analysis, clustering, classification, stock market prediction

VII. INTRODUCTION

The stock market is an important part of the economy of a country. The stock market plays a play a pivotal role in the growth of the industry and commerce of the country that eventually affects the economy of the country to a great extent. That is the reason that the government, industry and even the central banks of the country keep a close watch on the happenings of the stock market. The stock market is important for both the industry's point of view as well as the investor's point of view.

Whenever a company wants to raise funds for further expansion or settling a new business venture, they have to either take a loan from a financial organization or they have to issue shares through the stock market. In fact, the stock market is the primary source for any company to raise funds for business expansions. If a company wants to raise some capital for the business, it can issue shares of the company that is basically part ownership of the company. To issue shares for the investors to invest in the stocks a company needs to get listed to a stock exchange and through the primary market of the stock exchange they can issue the shares and get the funds for business requirements.

Stock price movements are driven by publications of financial news [1]. Investors' or sellers' decisions are made based on the available information considering how a stock market will go up or down for next few days. News articles incorporate information about a firm's fundamentals, the activities in which a firm is involved and the expectations of other market participants about future price changes [1]. An enormous amount of textual information is provided by real-time trading applications with a tremendously increased broadcasting speed. According to the reviewed literature, a predefined criterion is generally employed by researchers to select news articles relevant to an analyzed stock from a large collection of documents. After relevant articles are selected, the researchers tend to treat every article as having potentially the same impact on stock prices. Therefore, analyzing the behaviour and

performance of financial markets has become a major research field.

In this research very careful to select the relevant and reliable news articles and calculate the sentimental score for the influence feature. The influence feature is combined with other price features and input into Recurrent Neural Network (RNN) to predict the Sri Lankan stock market price.

The remainder of the paper is organized as follows Section II describes a review of research work related to this study. Section III illustrates the data collection process and the features proposed and Finally, conclusions are presented in Section IV.

VIII. LITERATURE SURVEY

Different datasets and techniques have been used for different researches to predict the stock market movement. Some of the work is defined in this section.

Muhammad Waqar, Hassan Dawood, Muhammad, Bilal Shahnawaz, Mustansar Ali Ghazanfar, Ping Guo [2] published "Prediction of Stock Market by Principal Component Analysis". In this paper, the problem of the high dimensionality of the stock exchange is investigated to predict the market trends by applying the (Principal Component Analysis) PCA with linear regression. PCA can help to improve the predictive performance of machine learning methods while reducing the redundancy among the data. Experiments are carried out on a high dimensional spectral of 3 stock exchanges. The accuracy of the linear regression classification model is compared before and after applying the PCA. The experiments show that PCA can improve the performance of machine learning in general if and only if relative correlation among input features is investigated and careful selection is done while choosing principal components. RMSE is used as an evaluation metric to evaluate the classification model [2].

Mustain Billah, Sajjad Waheed, Abu Hanifa published a paper regarding Stock Market Prediction Using an Improved Training Algorithm of Neural Network [3]. In this paper, an improved (Levenberg Marquardt) LM algorithm (Artificial Neural Network) ANN has been proposed. Then, this improved algorithm has been applied to stock market closing price prediction. ANN uses this improved algorithm for training, it shows 53% more accuracy in stock prediction than Adaptive Neuro Fuzzy Inference System (ANFIS). It also requires less memory allocation and computing time. This improved LM training algorithm proves neural network to be better computing tool for predicting closing stock price in Bangladesh Stock Exchange perspective [3].

A.J.P. Samarawickrama, T.G.I. Fernando [4] developed models to predict daily stock prices of selected listed companies of Colombo Stock Exchange (CSE) based on RNN Approach and to measure the accuracy of the models developed and identify the shortcomings of the models if present. SRNN, GRU and LSTM architectures were employed in building models. only the number of hidden neurons was changed while fixing the number of input variables to find the best model. In this study, when considering the forecast error (or the test error) MLP models produce the highest and the lowest errors. The forecasting accuracy of the best feedforward networks is approximately 99%. SRNN and LSTM networks generally produce lower errors compared with feedforward networks but in some occasions, the error is higher than feedforward networks. Compared to the other two networks, GRU networks are producing comparatively higher forecasting errors. MLP models produce the best results. This is because; in this study data for only the past two days were selected for inputs. If the number of past days considered for input variable selection was increased, RNN models would produce best results. They only limited to three companies in manufacturing, diversified and banking sectors. This study predicts only the closing prices, but models can be developed to predict open, low, high prices, transaction volumes and return on securities, etc [4].

Jiahong Li, Hui Bu, Junjie Wu developed a LSTM model which combined market information and investor sentiment to predict CSI300 index values in the Chinese stock market [5]. First, they deploy a Naïve Bayes sentiment classifier to assign all posts on stock forums to three classes: positive, negative, and neutral. And then they generate sentiment time series for subsequent work. Finally, they develop a deep neural network model which consists of a Long Short-Term Memory layer, a merged layer, a ReLU linear layer and a SoftMax layer. Trained on 90% of the entire data set, this model gives a prediction accuracy of 87.86% in the rest 10% of testing data, outperforming other input permutations and SVM method by at least 6% [5].

Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche did a research study regarding "Predicting Stock Price Movements Based on Different Categories of News Articles" [6]. This research study explores whether the simultaneous use of different financial news categories can provide an advantage in financial prediction system based on news. Five categories of news articles were considered: news relevant to a target stock and news relevant to its sub-industry, industry, group industry and sector. Each category of news articles was pre-processed independently, and five different subsets of data were constructed. The MKL approach was used for learning from different news categories; independent kernels were employed to learn from each subset. A number of different 709 types of kernels and kernel combinations were used. The findings have shown that the highest prediction accuracy and return per trade were achieved for MKL when all five categories of news were utilized with two separate kernels of the polynomial and Gaussian types used for each news category. The highest

kernel weights were assigned to the polynomial kernels indicating that this kernel type contributes the most to the final decision. The SVM and kNN methods based on a single category of news, either stock-specific (SS), sub-industry-specific (SIS), industry-specific(IS), group-industry-specific(GIS) or sector-specific(SeS), demonstrated worse performance than MKL. These results indicate that dividing news items into different categories based on their relevance to the target stock and using separate kernels for learning from these categories allows the system to learn and utilize more information about the future price behaviour which gives an advantage for more accurate predictions [6].

R. Batra and S. M. Daudpota did a research study regarding "Integrating StockTwits with sentiment analysis for better prediction of stock price movement" [7]. This research study exploits the benefits of sentiment analysis in the stock market industry. This paper has performed sentiment analysis on tweets related to Apple products, which are extracted from StockTwits (a social networking site) from 2010 to 2017. Along with tweets and used market index data which is extracted from Yahoo Finance for the same period. The sentiment score of a tweet is calculated by the sentiment analysis of tweets through SVM. Then sentiment score and market data are used to build a SVM model to predict the next day's stock movement. Results show that there is a positive relationship between people opinion and market data and proposed work has 75.22% training accuracy and 76.68% test accuracy in stock prediction. And it used NLP for the text prepossessing.

W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee did the research regarding "Stock market prediction using neural network through news on online social networks" [8]. In this paper, implement a model based on Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU) to predict the stock volatility in the Chinese stock market. Apart from that, select official accounts from Chinese largest online social networks - Sina Weibo and extract the content posted by these accounts to analyze the public moods.

Deepak Kumar Mahto and Lisha Singh published "A Dive into Web Scraper World" [9]. In this paper, used a web crawler to do the web scraping and python for implementation.

Shreya Upadhyay, Vishal Pant, Shivansh Bhasin and Mahantesh K Pattanshetti published a paper for Web Scraper for Massive Data Extraction [10]. This paper provides an insight into issues relevant to constructing a web scraper and concludes by describing the implementation of a web scraper for harvesting learning objects for an eLearning application [10].

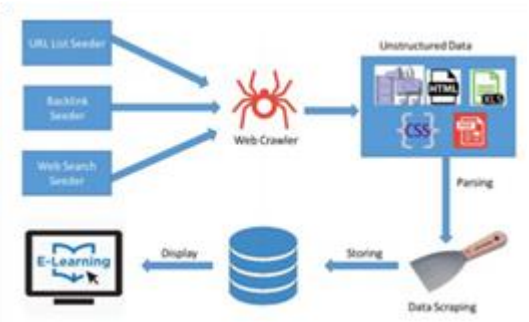


Fig 1: Architecture of a Web Scraper [10]

Xianlin Ma published "The Diseases Clustering for Multi-Source Medical Sets" [11]. In this paper, succeeded in clustering the diseases under multi-source medical sets. From using the Natural Language Processing technologies and k-center algorithm, all the disease in different sets was clustered into 3,633 clusters. And the diseases within the same cluster were described in many aspects, such as the cosine similarity, the proportion of the disease and so on. According to different clustering results, made a simple analysis [11].

Jaydeep Jalindar Patil and Nagaraju Bogiri [12] developed the system provides text categorization of Marathi documents by using the LINGO [Label Induction Grouping] algorithm. The system uses the dataset which contains 200 documents in 20 different categories. The result represents that for Marathi text documents LINGO clustering algorithm is efficient [12].

S. B. Bhaskoro, S. Akbar, and S. H. Supangkat, published "Extracting important sentences for public health surveillance information from Indonesian medical articles" [13]. The text summary method used in this research is a supervised learning approach. The sentence summary will be viewed as a sentence classification problem. Sentences in the article will be divided into three important sentences, such as the class of core sentences, supporting sentences and sentences of conclusion. Comparison to the test results performed includes feature selection consisting of title and noun, weighting consisting of TFIDF and TF and parameter values maximal marginal relevance consisting of 0.4, 0.6, 0.7, 0.8 while classification of sentence categories in this study using multinomial naïve Bayes method. The results obtained by performing a summary technique using feature selection title, weighting TF-IDF and MMR parameter value 0.7 to classify sentence categories get the results as follows 85.8% and 83.7% for calculation of precision and recall [13].

IX. METHODOLOGY

In this paper, extracted the news article from websites through NLTK, processed them from Natural Language

Processing (NLP) and sentiment analysis. After applied SVM in order to predict the sentiment of each article. After predicting sentiment, extracted historical data from the Colombo Stock Exchange (CSE). Developed a model for stock market prediction using stock price data and sentimental value to predict the change in sector-wise Srilankan stock market.

Fig 2 shows the entire workflow of the system. It mainly composed of 4 main parts which work integrating with one another. The database act as the central storage for storing the data all the data been backup to it. The overall system has 4 main functionalities were carried out under subparts of the system as shown in the diagram. The successful completion of the subparts can full fill the main functionalities which lead to the completion of the complete system. Those 4 functionalities are,

- Extracting unstructured data
- Algorithm for format the data
- Algorithm for sentiment analysis for each factor/sector.
- Algorithm for prediction

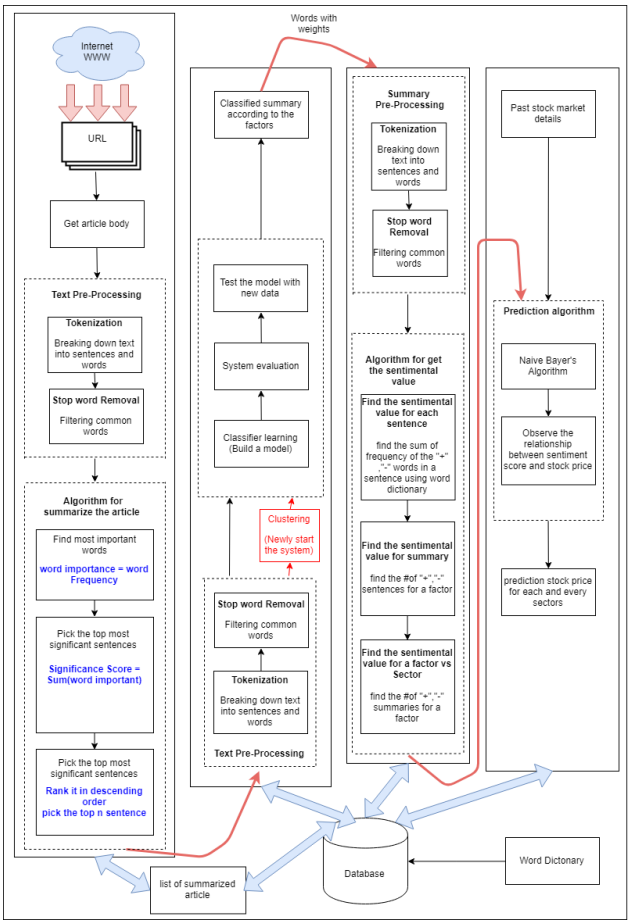


Fig 2: Architecture Diagram

A. Extracting Unstructured data

Data source are newspaper articles, blogs etc. Select reliable website and get the URL of the webpage as input and after some pre-processing, it will give output as a set of summaries of the articles.

In the newspaper articles are unstructured data. Therefore, the news website contains a lot of current article with big paragraphs. Because of that generate auto-summarize for a long piece of the article using a NLP program maintain the main story of every article to be used for the sentiment analysis. This step takes to obtain important sentences is to do a summary text that minimizes the number of paragraphs and minimize the sentence in the article to perform classify article category and calculate the sentimental value for the factors.

In auto-summarize is basically extract some important sentences from the text itself, and combine them together, to form the abstract. The idea being that some well-chosen sentences from the complete text together, can bring out the summary without having to read the entire text. For this use a rule-based approach to perform this abstract extraction. Those steps are,

1. Find the most important words (in the text)

   Word Importance = Word Frequency

2. Compute a significance score for sentences based on words they contain.

   Significance Score = Sum (Word Importance)

3. Pick the top most significant sentences (these will form the abstract)

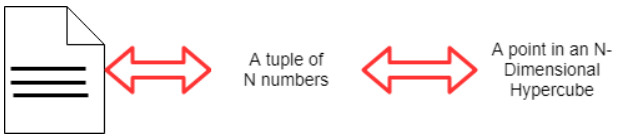
   Rank the sentences in descending order according to the significance score, then pick the top n sentences.

Fig 3: High level diagram for abstract extraction

The auto-summarize, at the high level there will be three important steps involved, first retrieve text that wants to summarize, from a web page. Which will involve downloading using Urllib2 and parsing the text using BeautifulSoup from the website.

Next process the raw text that is downloaded from a web page, by doing some preprocessing steps like

tokenizing text into sentences, tokenize sentences into words and remove stopwords.



Finally, apply the rule-based algorithm to identify the most important sentences in the article.

Fig 4: K-Means Clustering

B. *Algorithm for Format the data*

Take all the summaries and convert text to TF-IDF representation. In here each summary is a tuple of N numbers. Then it will do the clustering using the K-Means algorithm.

Find the most frequent words within each cluster and according to these words, can assign the underlying them (factors) for each cluster and do the classification using K-NN algorithm for upcoming extracted summaries.

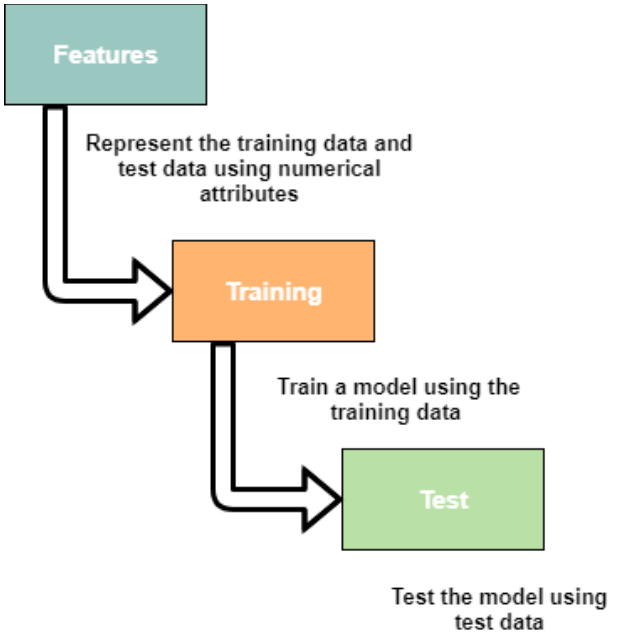


Fig 5: Classification setup

C. *Algorithm for sentiment analysis for each factor/sector*

Sentimental data for the factors will get from analyzing the article and compare with a created dictionary with the positive, negative words. Appling the SVM to predict the sentiment of each sentence of the summary and calculate the sentimental value for every factor [7].

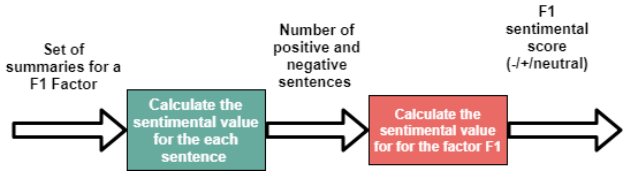


Fig 6: Sentimental analysis flow

D. *Algorithm for prediction*

Mainly using the Naive Bayes Algorithm But also for the accuracy testing purpose use RNN and Random Forest Algorithm. In Naive Bayes algorithm for the testing data, it gets more accurate predictive value than KNN algorithm. [14] Purpose of Naive Bayes algorithm is to find the minima/maxima of a function by iteratively moving in the direction of the negative slope of the function that it want to minimize/maximize. And also, it uses the supervised learning. 'Naive Bayes ', it is the fastest relative to other classification algorithms. It works on Bayes theorem of probability to predict the class of unknown data set. Naive Bayes model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

In Prediction Part, after sentimental analysis, the sentiment value will come to this section. And prediction will happen using a prediction algorithm. In that case, it will consider the factors along with the Sectors and finally return the output as a most accurate prediction of sector-wise Stock Market Price.

Table 1: Sample training data for prediction

Date	F1	F2	F3	F4	S1 Val	Status
7/1/18	6	-2	10	2	1500	Up
7/2/18	5	-10	2	-2	1000	Down
7/3/18	6	-3	8	2	1300	Up
7/4/18	6	0	0	6	1600	Up
7/5/18	0	0	0	0	1600	----
7/6/18	2	0	0	0	1608	Up
7/7/18	0	2	0	0	1610	Up

Table 2: Sample testing data for prediction

Date	F1	F2	F3	F4	S1 Val	Status
7/8/18	2	0	2	5	?	?

X. CONCLUSION

Early research on Sri Lankan stock market prediction was totally based on a random walk and numerical prediction but with the introduction of real-time behavioural finance, get the sentimental score of the abstract extraction of news article will consider while predicting the stock movement. Making it more efficient it uses the idea of sentimental analysis of the factor wise summary of the article through machine learning models.

This paper gives the idea by collecting factor related sentiment data from websites, build the SVM model to calculate the sentimental value, and predict the Sector wise stock market value using Naïve Bayes algorithm with more accuracy. If increase the size of the data set will gives more accuracy.

REFERENCES

[22] Y. Shynkevich, T. M. McGinnity, S. Coleman, and A. Belatreche, 'Predicting Stock Price Movements Based on Different Categories of News Articles', in 2015 IEEE Symposium Series on Computational Intelligence, 2015.

[23] M. Waqar, H. Dawood, P. Guo, M. B. Shahnawaz, and M. A. Ghazanfar, 'Prediction of Stock Market by Principal Component Analysis', in 2017 13th International Conference on Computational Intelligence and Security (CIS), 2017

[24] M. Billah, S. Waheed, and A. Hanifa, 'Stock market prediction using an improved training algorithm of neural network', in 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), 2016.

[25] A. J. P. Samarawickrama and T. G. I. Fernando, 'A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market', in 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), 2017.

[26] Jiahong Li, Hui Bu, and Junjie Wu, 'Sentiment-aware stock market prediction: A deep learning method', in 2017 International Conference on Service Systems and Service Management, 2017.

[27] Y. Shynkevich, T. M. McGinnity, S. Coleman, and A. Belatreche, 'Predicting Stock Price Movements Based on Different Categories of News Articles', in 2015 IEEE Symposium Series on Computational Intelligence, 2015.

[28] R. Batra and S. M. Daudpota, "Integrating StockTwits with sentiment analysis for better prediction of stock price movement," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018.

[29] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Stock market prediction using neural network through news on online social networks," in 2017 International Smart Cities Conference (ISC2), 2017.

[30] Deepak Kumar Mahto and Lisha Singh, "A dive into Web Scraper world", in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.

[31] S. Upadhyay, V. Pant, S. Bhasin, and M. K. Pattanshetti, 'Articulating the construction of a web scraper for massive data extraction', in 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017.

[32] L. Li, S. Xu, S. Wang, and X. Ma, 'The Diseases Clustering for Multi-source Medical Sets', in 2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI), 2016.

[33] J. J. Patil and N. Bogiri, 'Automatic text categorization: Marathi documents', in 2015 International Conference on Energy Systems and Applications, 2015.

[34] S. B. Bhaskoro, S. Akbar, and S. H. Supangkat, "Extracting important sentences for public health surveillance information from Indonesian medical articles," in 2017 International Conference on ICT For Smart Society (ICISS), 2017.

[35] Sunil Ray, "6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)", Analytics Vidhya, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Accessed: 29- Mar- 2018].