Analytics of stock market prices based on machine learning algorithms

Puteri Hasya Damia Abd Samad[1], Sofianita Mutalib[2], Shuzlina Abdul-Rahman[3]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia

Article Info

Article history:

Received Jan 29, 2019
Revised Mar 2, 2019
Accepted May 13, 2019

Keywords:

Bursa Malaysia
Frequent itemset
Machine learning
Stock market prices prediction
Text mining

ABSTRACT

This study focuses on the use of machine learning algorithms to analyse financial news on stock market prices. Stock market prediction is a challenging task because the market is known to be very volatile and dynamic. Investors face these kinds of problems as they do not properly understand which stock product to subscribe or when to sell the product with an optimum profit. Analyzing the information individually or manually is a tedious task as many aspects have to be considered. Five different companies from Bursa Malaysia namely CIMB, Sime Darby, Axiata, Maybank and Petronas were chosen in this study. Two sets of experiments were performed based on different data types. The first experiment employs textual data involving 6368 articles, extracted from financial news that have been classified into positive or negative using Support Vector Machine (SVM) algorithm. Bags of words and bags of combination words through Apriori algorithm are extracted as the features for the first experiment. The second experiment employs the numeric data type extracted from historical data involving 5321 records to predict whether the stock price is going up (positive) or down (negative) using Random Forest algorithm. The Rain Forest algorithm gives better accuracy in comparison with SVM algorithm with 99% and 68% accuracy respectively. The results demonstrate the complexities of the textual-based data and demand better feature extraction technique.

Corresponding Author:

Sofianita Mutalib,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia.
Email: sofi@tmsk.uitm.edu.my

1. INTRODUCTION

Stock market investors normally have to make difficult decisions based on the assumptions that the presumed price is different from the current stock market price due to its intrinsic value [1-2]. The intrinsic value of the stock is considered as constant within a short period of time because opinions or decisions that have been made by investors are not expected to change drastically in a short period of time. Investors make comparisons between the perceived intrinsic value and its market value, and later investors make decisions on buying or selling or holding based on the current situation [2-3]. Billions of money are traded or exchanged every day whereby the investors are hoping that they will make profit instead of losses. Behavior of investors can affect stock prices, and investors influence stock prices by using information that are available in the public domain to predict the results of how the market will react [4-5]. This makes stock market analytics an extremely interesting area and its development is worthwhile for the investors. Market prediction that is effective might help the investors in terms of trading advice or as a key component for stockbrokers. Furthermore, prediction models can help investors in providing helpful information like market direction in the future[5-7]. For example, if the selected stock is predicted to increase, buying a stock during that time may increase profits.

Moreover, financial news article can play a huge role in influencing the movement of stocks [3, 4, 7]. Furthermore, news normally carry valuable information about a particular firm foundations and also expectations of related market participants. However, the investors could be facing problems in analysing many articles [5] and they are overwhelmed by information and in the end, it is still a hard decision to make to buy or to sell in order to get optimum profit. The investors are aware that the news and analyst reports contain rich information about stock markets [7-9]. Hence, analyzing news about particular stock, including online feeds from huge sources is an important resource in stock market prediction [6-7]. Analyzing the information individually or manually is difficult [7-9]. Thus, we need an appropriate tool with suitable algorithms to solve this kind of problem, and the analytics would be carried out in an efficient and proper way through text mining. Many methods have been proposed to predict the directions of stock market prices but with numerous attempts in trying to get accurate stock price effects such as positive or negative, many of them barely exceeded 58% [7, 10-12]. From the observation, it proves that the data should be duly analyzed in order to get the desired output. Supply and demand can affect stock prices changes of a particular stock and changes in the market are associated with the sentiments of investors. Therefore, news articles may carry timely information about a particular stock market. Other difficulties in analyzing the stock market from news are that it might not be entirely predictable [3-11], and textual data has its complexities of modelling the market, which is dynamics [5, 13, 14]. Since information increases day by day, it is quite challenging for some investors to consider all the available information. Therefore, a textual classification engine that can automatically process the textual data from financial news has become more relevant and important.

This paper presents the “Analytics of Stock Market based on Machine Learning Algorithms” by using financial news and historical price data. The study in this paper focuses on understanding the use of machine learning algorithms in predicting stock prices on the FTSE Bursa Malaysia KLCI (FBMKLCI). The problem in predicting stock market prices is resolved through data mining with machine learning algorithms such as the Support Vector Machine (SVM) and Random Forrest (RF). The remainder of this paper is organized as follows: The next section i.e. Section 2 discusses the related work of machine learning algorithms while Section 3 describes the research methodology for the study. Section 4 presents the analysis and results of the experiments. Finally, Section 5 concludes the research with the recommendations for future work.

2. RELATED STUDIES

As the name implies predictive analytics refers to techniques used to predict future or unknown events. It uses various techniques ranging which include data mining, machine learning, natural language processing, and statistics. The procedures involved include analyzing existing data in order to make predictions of the future. It uses predictive modelling and analytical techniques to carry together the information from business processes, information management tool and modelling algorithms in order to produce outcomes about the future stock prices. There are two approaches commonly adapted by market professionals to predict stock market prices namely, (1) "chartist" or "technical" theories and (2) the theory of fundamental or intrinsic value analysis [2], [15]. Based on Chartist theories [15], the past behavior of stock market prices have large amounts of information that will lead to its future behavior. The “pattern” of past behavior will repeat and tend to happen again in the future. By analyzing price charts, it can help develop a clearer understanding of the patterns and the valuable patterns observed can then be used to predict the future behavior of prices and can help increase expected gains. As this research is of the opinion that data can reveal patterns and aid understanding, we therefore will make use of the existing news and adapt available modelling for the classification and prediction. The goal of the stock market prediction is basically about making the prediction of some aspects of the stock market such increase or decrease in prices [6]. The incoming news can be analysed and classified as trying to portray messages that stock price is going to increase or decrease.

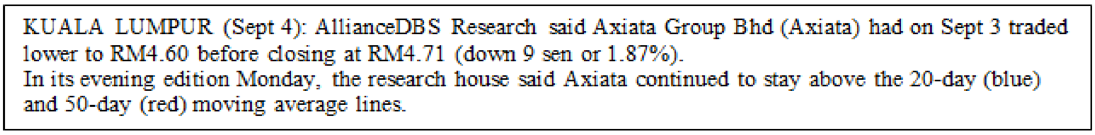
In general, in stock market analytics, there are descriptive and predictive measures [7, 17]. Descriptive measures involves calculations of simple measures of composition and the distribution of variables. Fundamental and technical analyses include predictive measures in the analysis. Fundamental analysis focuses on studying the competitor, markets, and business [7, 11]. Technical analysis focuses on historical prices analysis, which leads to determining the upcoming results of stock prices [11, 15, 18]. Text mining is a technique that involves extracting valuable information and pattern [12-14]. There is a huge amount of accessible data in the information production that would be the resources for the fundamental analysis. A bigger amount of text is spread over the internet like blogs or social media. It is quite challenging to determine the patterns and trends to extract valuable knowledge; nevertheless, it is possible to analyze this huge amount of data and remove unnecessary information from it. Next, this kind of information can be used in many applications such as market investigation and stock market performance forecasting [7, 8, 11, 12]. Text mining fulfills its main goal by the possible identification of useful, accurate, understandable correlations and patterns in the data, and this can be achieved by modelling using descriptive and predictive nature [6-17]. Descriptive model is basically about finding and identifying the patterns or relationships in datasets, while predictive model is about making a prediction about what is coming next using meaningful existing data. Categorization technique in predictive analytics involves using supervised learning method. It requires the examples of desired output to classify new documents such as Support Vector Machine (SVM), Support Vector Regression (SVR), Decision Tree (DT), Artificial Neural Network (ANN), and Naïve Bayes (NB), all of which can be used to categorize the text or documents [6, 5, 10, 17]. This process includes pre-processing, indexing, dimensional reduction, and classification [13, 19-21]. According to Gaikward et al., [12] information or textual data can be presented in four forms namely Term Based Method (TBM), Phrase Based Method (PBM), Concept Based Method (CBM) and Pattern Taxonomy Method (PTM). Most of the researches in stock market using bags of word (BoW), which is the TBM to represent the features [5, 7, 10, 18]. Though so there are other studies which have used combination of words as the feature set [22-23], and this representation is hardly found in stock market financial news. This effort in text mining have been employed to enhance the relevance and accuracy of results [24-28].

3. RESEARCH METHOD

Preliminary studies take place at the beginning of the process of the development. During this phase, all ideas and information are gathered through reading the relevant journals, articles, books and authoritative websites on stock market analyses. This study attempts to capture data regarding the stock market on online websites i.e. The Edge, (https://www.theedgemarkets.com, from 2014 till 2018) with an example in Figure 1 for an article dated September 4, 2018, with the title “Hurdle for Axiata at RM4.83, says AllianceDBS Research”. This study aims to narrow down the scope and have chosen only five of the top market constituents from FTSE Bursa Malaysia Kuala Lumpur Composite Index (FBMKLCI). Meanwhile, the historical prices were captured from Yahoo! Finance. The related companies are Axiata Group, CIMB Group Holdings, Malayan Banking, Petronas Chemicals Group and Sime Darby.

3.1. Numerical Data

Numerical data are structured information, and in this study, it refers to the historical prices that were gathered from Yahoo Finance website. The information linked with the historical prices are also captured that includes the set of variables namely open price, high price, low price, adjusted close and volume of trade. The historical prices are used to allow the labelling or tagging the articles to their ability to affect stock prices in general.



KUALA LUMPUR (Sept 4): AllianceDBS Research said Axiata Group Bhd (Axiata) had on Sept 3 traded lower to RM4.60 before closing at RM4.71 (down 9 sen or 1.87%).
In its evening edition Monday, the research house said Axiata continued to stay above the 20-day (blue) and 50-day (red) moving average lines.

Figure 1. Example of textual data from the Edge market publication

3.2. Textual Data

As opposed to numerical data, textual data is basically in the form of text which is unstructured data on the online news. A script is incorporated and crawled for data related to the stock market, and in our study, this is based on chosen companies in Malaysia found on The Edge Market website. Next, the gathered articles which contain useful information such as the title, article, and time the news was issued. These articles are stored in JSON file format for further analysis using statistical tools for simple observations. Data analysis was carried out to comprehend the distribution of the textual dataset that is captured. Furthermore, the process also helped to discover the noises or outliers in the data. From the observations done, the redundancies of the articles can be avoided including unrelated articles in other languages such as Mandarin or Indian articles. Next, textual format, which is unstructured format, needs to be transformed to structured format in order to carry out further training and testing processes. The subsequent step is data cleansing to remove the redundancies of the articles or documents by finding the similarities of these type of unstructured data. Python script was used to find the similarity of the current article with next article, and the unnecessary articles were removed. After removing the redundant articles, the punctuation and symbols were discarded from the content (of the articles). Table 1 shows that there are no articles captured on the particular dates, it will be marked as 'None' and will removed.

Table 1. Sample of Historical Prices and Related Articles

Date	Article	Open	High	Low	Close	Adj Close	Volume
2/1/2014	Article 1	6.92	6.94	6.84	6.92	6.176871	4566100
3/1/2014	None	6.88	6.88	6.80	6.80	6.069758	4261800
6/1/2014	None	6.80	6.81	6.77	6.80	6.069758	10519600
7/1/2014	Article 2	6.80	6.80	6.76	6.79	6.060832	6166400
8/1/2014	Article 3	6.79	6.79	6.76	6.77	6.042980	4393300

3.3. Bags of Word (BoW)

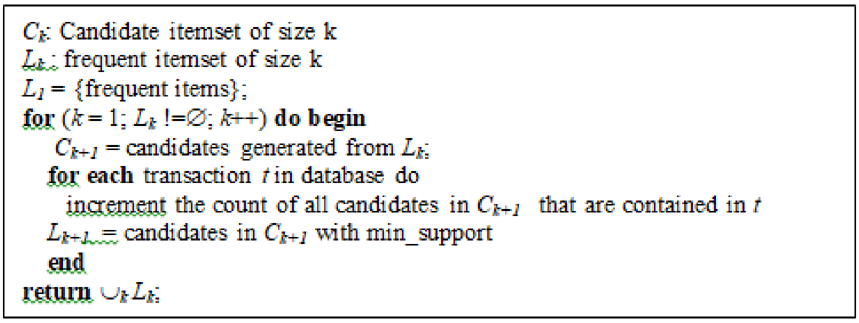
The bags of word (BoW) were extracted through several steps on the articles, including tokenization, stop words removal, Porter's word stemming and also removal of the attributes (words) that carry irrelevant information. After removing the unnecessary attributes (words), the Chi-square test was performed on each word to reduce the BoW into significant and meaningful features that represent explanatory power of the features. Smaller dataset is useful for learning process as it will reduce computational time throughout the development. The words with zero Chi-square value are the words that appear consistently in positive and negative documents. On the contrary, the features that are not consistent in positive and negative tend to get higher values in Chi-square and the highest value. The total number of BoW for positive and negative words are 7121 and 6878 respectively. Beyond that, the frequency of each word was calculated. The first 100 words that held the highest frequency were selected to be used in the training and testing datasets. Sample of the outputs from tokenization and stemming can be found in Figure 2.

Figure 2. Output of tokenization and stemming process

Original Words	Shares valued at RM209.74 million				
Tokenization	Share	valued	at	RM209.74	million
Stemming	Share	value	at	RM209.74	million

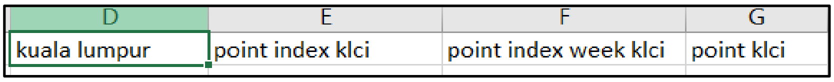
3.4. Bags of Combination of Words (BoCW)

The second set of features were constructed using the bags of combination of words (BoCW). The BoCW was extracted by applying Apriori algorithm using WEKA, as shown in Figure 3 [22]. The generated rules from Apriori was set with minimum support of 0.6 and confidence level is equal to 1. The top 1000 rules were selected to represent the features in this new dataset. Figure 4 displays the example of BoCW in the dataset. BoCW was mapped back to each of the article to get the frequency of each feature.



```
$C_k$: Candidate itemset of size k
$L_k$: frequent itemset of size k
$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
   for each transaction $t$ in database do
     increment the count of all candidates in $C_{k+1}$ that are contained in $t$
   $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
   end
return $\cup_k L_k$;
```

Figure 3. Pseudo code for apriori algorithm



D	E	F	G
kuala lumpur	point index klci	point index week klci	point klci

Figure 4. Example of BoCW

3.5. Data Labeling

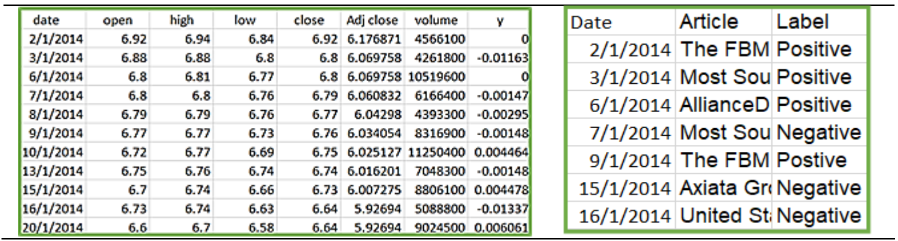
Data labelling works by calculating the rate of change based on the opening and closing prices from the historical prices according to the label of positive, where closing price is higher than opening or negative, where closing price is lower than opening [11]-[15]. Data labelling is meant for textual data for determining the articles as to whether they give positive or negative outputs to the stock as shown in Table 2. The equation is as:

y = positive (1)

negative (2)

$$r = \frac{Close - Open}{Open}$$

Table 2. Labelling the Historical and News



Historical information

date	open	high	low	close	Adj close	volume	y
2/1/2014	6.92	6.94	6.84	6.92	6.176871	4566100	0
3/1/2014	6.88	6.88	6.8	6.8	6.069758	4261800	-0.01163
6/1/2014	6.8	6.81	6.77	6.8	6.069758	10519600	0
7/1/2014	6.8	6.8	6.76	6.79	6.060832	6166400	-0.00147
8/1/2014	6.79	6.79	6.76	6.77	6.04298	4393300	-0.00295
9/1/2014	6.77	6.77	6.73	6.76	6.034054	8316900	-0.00148
10/1/2014	6.72	6.77	6.69	6.75	6.025127	11250400	0.004464
13/1/2014	6.75	6.76	6.74	6.74	6.016201	7048300	-0.00148
15/1/2014	6.7	6.74	6.66	6.73	6.007275	8806100	0.004478
16/1/2014	6.73	6.74	6.63	6.64	5.92694	5088800	-0.01337
20/1/2014	6.6	6.7	6.58	6.64	5.92694	9024500	0.006061

News

Date	Article	Label
2/1/2014	The FBM	Positive
3/1/2014	Most Sou	Positive
6/1/2014	AllianceD	Positive
7/1/2014	Most Sou	Negative
9/1/2014	The FBM	Postive
15/1/2014	Axiata Gr	Negative
16/1/2014	United St	Negative

For each of the data collected i.e. textual and numerical, two separate models were developed. The Support Vector Machine (SVM) was used to learn from textual data, and the Random Forrest was used to learn from numerical historical data. The entire processes involved in the research is provided in Figure 5. Two sets of experiments was performed as described in the next section

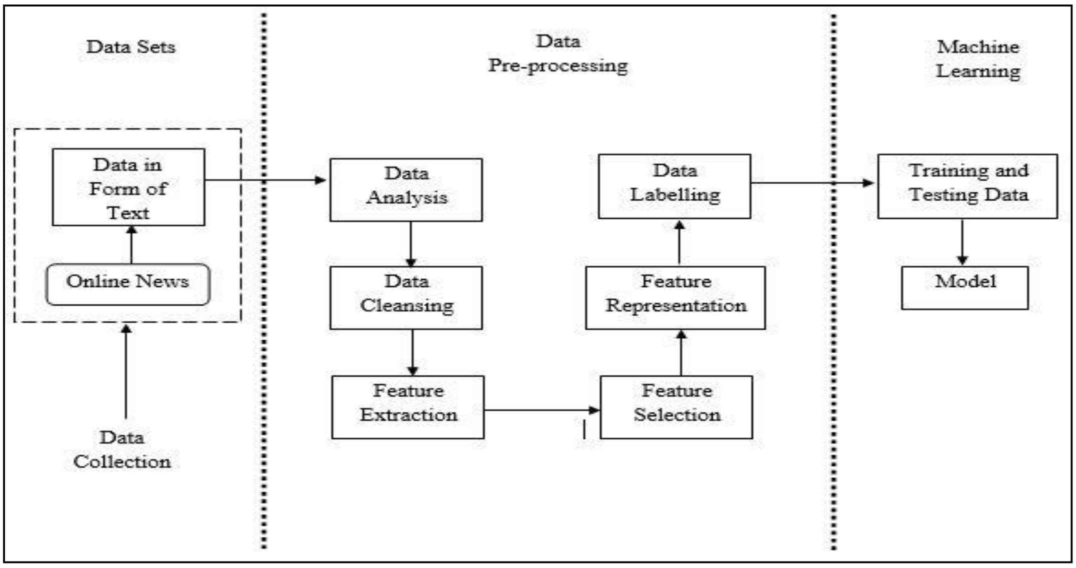


Figure 5. Flow of the processes in the research

3.6. Experiment 1: Classification Positive or Negative based on BoW and BCoW

The process of modelling the prediction system with SVM algorithm was performed by splitting data into training and testing sets. The optimum parameters for the model were obtained with a series of experiments during model testing. Classifier accuracy on a test set is measured by its ability to classify them accurately.

3.7. Experiment 2: Prediction using Numeric Values

This second experiment applied the Random Forrest model to produce price prediction based on historical prices. The (Random Forrest) model is efficient at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. The numerical dataset is divided into training and testing according to the percentage of 70%/30% and 80%/20% respectively.

4. RESULTS AND ANALYSIS

4.1. Analysis of Textual Data

Data analysis on the dataset was carried out to understand the data and presented graphically to comprehend its distribution. Figure 6 illustrates the total data including textual data and historical prices from each company. It shows that the total number of textual and historical data for each company is totally different. As such, the articles without historical data were removed. Table 3 shows the results after removing all redundant articles from the news. The total amount of articles that have been discarded is 523 from 6368 articles. Nevertheless, the remaining dataset, which contains 5845 articles can be used functionally in this study.

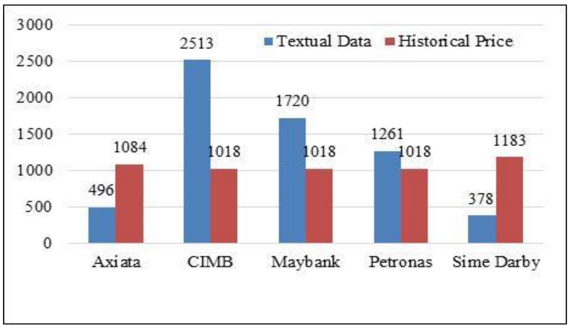


Figure 6. The total number of data from each company

Table 3. Total Number of Articles After Removal and Cleaning Process

Company	Raw Article	Removed Data	Cleaned Article
Maybank	1720	154	1565
Axiata	496	22	474
CIMB	2513	182	2330
Petronas	1261	143	1114
Sime Darby	378	22	355

4.2. BoCW as Feature from Textual Data

This study only takes the first 20 combination of words that have confidence level equal to 1 for each company. Table 4 shows the example of combinations of positive and negative words from Axiata.

Table 4. Combination of Words

List of Positive Words	List of Negative Words
1.develop project	1.score number fundament
2.gain index	2.revenu busi dividend say
3.high index	3.valuat number fundament
4.buy index	4.oper dividend
5.index klci point	5.edg fundament
6.revenu project	6.score number fundament
7.sale project	7.close index
8.point oil	8.declin week
9.capital develop	9.fell klci
10.sale develop	10.declin index

4.3. SVM Modelling for Textual Data

SVM was employed to train and test the datasets and the evaluation is based on the accuracy of scores. The datasets are divided into training and testing according to the given percentage of 70%/30%, 80%/20%, and 90%/10% respectively. Scikit-learn SVM implementation was employed with several parameters for this study. The kernel that is used in this study is the Radial Basis Function (RBF). Next, this study compares the prediction between BoW and BoCW for each company. Each of it takes only a maximum features of 500. Table 5 shows the scoring between BoW and BoCW. As can been seen, the set of features that demonstrated the best accuracy is BoCW with training and testing percentage of 90% and 10%, whereby the average score of all companies is 68.49%, although there are companies where the BoW is well predicted compared to BoCW. However, the final accuracy for each company was approximately 60%, and it is consistent with the accuracies recorded by other researchers in related studies except for Sime Darby because the dataset is biased toward positive words. In comparison of BoW and BoCW, the BoCW model shows slightly higher accuracy from BoW. This could be an indication that BoCW is better at representing the data set for machine learning model.

Table 5. Accuracy for BoW and BoCW Feature

Company	BoW (Accuracy %)			BoCW (Accuracy %)			
	70/30	80/20	90/10	70/30	80/20	90/10	Average
Maybank	62.96	63.25	59.83	61.25	58.97	62.39	61.44
Axiata	61.86	60.76	50.00	58.47	56.96	62.50	58.43
CIMB	55.03	54.79	54.10	54.84	60.27	58.47	56.25
Petronas	55.13	59.43	60.23	58.93	55.43	59.09	58.04
Sime Darby	99.02	98.53	100	99.02	98.53	100	99.18
Average	66.80	67.35	64.83	66.50	66.03	68.49	

4.4. RF Prediction using Historical Data

RF was used to make a prediction based on historical prices. The trend analysis was carried out with information of actual price and prediction price by date and month, and it was found that the actual and prediction prices are quite similar because the accuracy of price prediction is high. An interesting trend line on the analysis shows that CIMB experienced decreasing stock market price over time. Therefore, investors might have to really think thoroughly to decide whether to invest in CIMB. Next, a dashboard of classification of positive and negative articles section and its prediction price was prepared. By showing the articles and its prediction price, the users can benefit from the information to buy or not the company's stock. It also gives the actual price of the stocks for users to see how accurate the prediction price is. Figure 7 shows the partial screenshot of the dashboard for the trend line graph for CIMB articles.

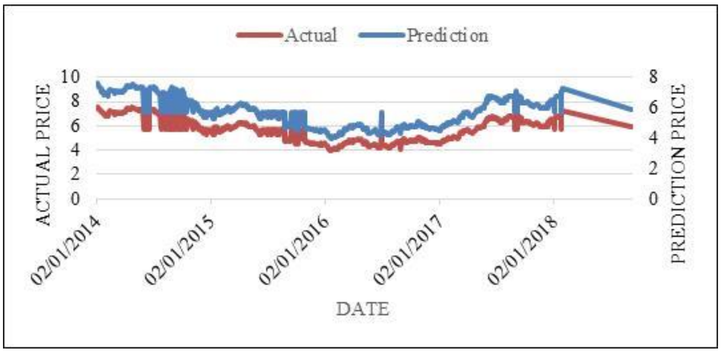


Figure 7. Trend lines of prediction vs actual

5. CONCLUSION

This paper presents the "Analytics of Stock Market (Prices) based on Machine Learning Algorithms" by using financial news and historical price data. The dataset was collected from the Edge Market publication for textual data and Yahoo! Finance for historical data. The problem in predicting the stock prices is modelled through data mining with machine learning algorithms namely the Support Vector Machine and the Random Forest. Two sets of experiments for the textual data, with BoW and BoCW and another set of experiment for the historical data by using the percentage split method was performed. The BoCW was extracted using Apriori algorithm with confidence equal to 1. The selection of words/combination of words is based on the frequency in the whole articles. Based on the results produced, the analytics gained more than 60% accuracy for the textual analytics, and 90% for the numerical analytics. With the use of text processing and machine learning, this study was able to deliver the appropriate methods and techniques used by the other researchers. This study also produced the recommendations such as display the future price of a particular company using dashboard, this information can be helping investors to decide in buying the product or not to invent. Furthermore, future works need to expand the companies selected to be experimented/studied in this study. This is because many investors can use this dashboard as reference to compare the prediction made by any other tools as well. Last but not least, the output of this study can also be the integrated in report to monitor the real time of stock market prediction.

ACKNOWLEDGEMENTS

The authors would like to thank the Research Management Centre of Universiti Teknologi MARA for supporting this research with LESTARI 111/2017 grant and Institute of Quality & Knowledge Advancement (InQKA) for the publication support.

REFERENCES

[1] R. Myšková, *et al*., “Predicting Abnormal Stock Return Volatility Using Textual,” *Journal of Amfiteatru Economic*, vol. 20(47), pp. 185-202, 2017.

[2] M. Khan, et al., “Financing and monitoring in an emerging economy: Can investment efficiency be increased?,” *Journal of China Economic Review,* vol. 45(C), pp 62-77, 2017.

[3] S. J. Grossman and J. E. Stiglitz, “On the impossibility of informationally efficient market,” *Journal of The American Economic Review*, vol. 70(3), pp. 393–408, 1980.

[4] B. Rosenberg, *et al*., “Persuasive evidence of market inefficiency,” *Journal of Portfolio Management,* vol. 11(3), pp. 9–16, 1985; DOI: https://doi.org/10.3905/jpm.1985.409007.

[5] R. Schumaker and H. Chen, “Textual analysis of stock market prediction using breaking financial news,” *Journal of ACM Transactions on Information Systems,* vol. 27, pp. 1-19, 2009.

[6] A. Nikfarjam, *et al*., “Text mining approaches for stock market prediction,” *The 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore,* vol. 4, pp. 256-260, 2010.

[7] M. Hagenau, *et al*., “Automated news reading: Stock price prediction based on financial news using context-specific features,” *Journal of Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1040-1049, 2012.

[8] E. Lupiani-Ruiz, *et al*., “Financial news semantic search engine,” *Journal of Expert Systems with Applications,* vol. 38, pp. 15565–15572, 2011.

[9] M. Khan and S. Khan, “Data and information visualization methods, and interactive mechanisms,” *International Journal of Computer Applications,* vol. 34(1), pp. 1–14, 2011.

[10] K. Lee and R. Timmons, “Predicting the stock market with news articles”, pp. 1-8, 2007.

[11] A. Rahman, et al., “Mining Textual Terms for Stock Market Prediction Analysis Using Financial News”, pp. 293–304, 2017.

[12] S. Gaikwad, et al., “Text Mining Methods and Techniques,” *International Journal of Computer Applications*, vol. 85(17), pp. 975–8887, https://doi.org/10.5120/14937-3507, 2014.

[13] T. Ramzan, *et al*., “Text Mining: Techniques, Applications and Issues,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7(11), pp. 414–418. 2016.

[14] W. Fan, *et al*., “Tapping the power of text mining,” *Journal of Communications of the ACM,* vol. 49(9), pp. 76–82, 2006

[15] E. Fama, “The Behavior of Stock Market Prices,” *Journal of Business,* vol. 38, no. 1, pp. 34-105, 1965, doi:10.1086/294743

[16] R. James, *et al*., 2012, “Business Analytics: The Next Frontier for Decision Sciences”, 2019, http://faculty.cbpp.uaa.alaska.edu/afef/business_analytics.htm.

[17] K. Agyapong, *et al*., “An Overview of Data Mining Models (Descriptive and Predictive),” *(IJournals) International Journal of Software & Hardware Research in Engineering,* vol. 4(5), pp. 53–60, 2016.

[18] M. Dang and D. Duong, “Improvement methods for stock market prediction using financial news articles,” *(NICS) 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science,* pp. 125–129, 2016.

[19] R. Agrawal and M. Batra, “A detailed study on text mining techniques,” (IJSCE) *International Journal of Soft Computing and Engineering,* vol. 2, no. 8, pp. 2231–2307, Jan 2013.

[20] Lam, *et al*., “Automatic Text Categorization and Its Application to Text Retrieval” *IEEE Transaction Knowledge and Data Engineering,* vol. 11(6), pp. 865–879, 1999.

[21] A. Khadjeh Nassirtoussi, et al., “Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment,” *Expert System Application [Internet],* vol. 42(1), pp. 306–324, 2015.

[22] R. Agrawal, *et al.*, "Mining association rules between sets of items in large databases," *ACM SIGMOD International Conference on Management of Data,* vol. 22(2), pp. 207-216, 1993.

[23] S. Mahmood, *et al.*, “Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets,” *The Scientific World Journal*, vol. 2014, 11 pages, 2014. https://doi.org/10.1155/2014/973750.

[24] Zhong, Li and Wu, “Effective pattern discovery for text mining,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24(1), pp. 30–44, 2012.

[25] C. P. Chen, and C. Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” *Information Sci- Ences,* vol. 275, pp. 314–347, 2014.

[26] S. Anwar and S. Sun, “Financial development, foreign investment and economic growth in Malaysia,” *Journal of Asian Economics*, vol. 22(4), pp. 335–342, 2011. https://doi.org/10.1016/j.asieco.2011.04.001

[27] M. Fikri and R. Sarno, “A comparative study of sentiment analysis using SVM and SentiWordNet,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13(3), pp. 902-909, 2019. http://doi.org/10.11591/ijeecs.v13.i3.pp1087-1094

[28] E. A. Abdullah, *et al*., “Modelling volatility of Kuala Lumpur composite index (KLCI) using SV and garch models,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13(3), pp. 1087-1094, 2019. http://doi.org/10.11591/ijeecs.v13.i3.pp1087-1094