Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Prediction of Malaysian stock market movement using sentiment analysis

To cite this article: Low Cheng Kuan *et al* 2019 *J. Phys.: Conf. Ser.* **1339** 012017

Journal of Physics: Conference Series **1339** (2019) 012017 doi:10.1088/1742-6596/1339/1/012017

Prediction of Malaysian stock market movement using sentiment analysis

[1]Low Cheng Kuan*, [1]Maizatul Akmar Ismail, [1]Tasnim M. A. Zayet, and [2]Shuhaida Mohamed Shuhidan

[1]Department of Information Systems, University of Malaya, Kuala Lumpur, Malaysia
[2]Accounting Research Institute (ARI) Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia

*wqd170013@siswa.um.edu.my

Abstract. Financial and business news contain various information about different companies, stock markets and other financial information. This information could be useful for predicting the stock market movement. The aim of this study is to determine whether financial news could be used to predict the Malaysian stock market movement. The sentiment analysis and classification were done using Hybrid Naïve Bayes algorithm. The data for this study was collected from Genting Berhad for a period of 11 months. The method resulted in news classification accuracy of 68.75% and showed a correlation of 58.41% between historical stock price and the sentiment data.

1. Introduction

In the past, people were expressing their opinions by talking, through interviews, writing articles and other traditional mean. In addition, people are getting news from newspaper, T.V news and other traditional sources. Since the emergence of the Internet, where web-based applications were created, social media became one of the application that created a huge change and remark in people's life. Through social media apps people can create a whole virtual community, add friends, share their trips details, share their opinions about services and products, share their preferred artists, movies and many more. Moreover, news starts to be published through social media, and becomes an important basis of real time data

Information published on social media are used in many fields like recommendation systems [1, 2], intelligent transportation systems [3, 4], politics [5] and others. In these fields, sentiment analysis is used to discover users' opinion, expectations, preferences, level of satisfaction and so on. In the field of economics, the news published on different applications allow traders to get real time information about finance and business [6], where it is found that there is a strong correlation between financial news and the stock market volatility [7]. One of an important use of social media news is predicting the stock market movements. The stock markets reflect the "moods" of people in the market. These moods are mostly affected by the financial news released regarding a company, or about an overall market [8]. People's opinion on social media like Twitter can predict the Dow Jones Industrial

Average (DJIA) value with 87.6% accuracy as well as reducing the Mean Average Percentage Error (MAPE) by more than 6% [9]. Dictionary-based method for analyzing the news headlines and press releases to get the sentiment indicators i.e. positive, neutral and negative are studied by Chowdhury et.al (2014) . They obtained 67% correlation between news sentiment and the closing stock prices of 15 individual companies as they got an average accuracy of 70.1% for identifying the correct sentiment.

Hybrid Naive Bayes (HNB) is another method used for sentiment analysis to predict the stock market. HNB merges two sentiment analysis methods, the lexical approach and Naive Bayes (NB) classifier [11-13]. NB is a supervised machine learning classifier, it is known for its fast performance, high accuracy and it requires a relatively small training data to train the model [14]. Furthermore, the lexical approach needs a shorter time in performing the sentiment analysis since it uses predefined dictionaries for the analysis. HNB is used to prove the EMH for Indonesian foreign exchange market [13]. Their work achieved a 69% of correct prediction for Twitter sentiments as it shows the Indonesian foreign exchange market has a positive correlation of 65.7% with the Twitter sentiment. Another similar study is [12] in which they used Twitter data to predict stock market using HNB. Their result is even better and more significant, reaching to 90.38% of classification accuracy. However, there's lack of work that uses news in predicting the sentiment of financial data. Thus, we chose to use HNB to predict the Malaysian stock market movements. This research is conducted to answer the following questions:

1. Is there any correlation exist between sentiment values from the news release and the stock price movement in Malaysia? (RQ1)
2. Can we trade Malaysian companies' stocks based on the sentiment data we get from the news released? (RQ2)

This paper will be organized as follow: the second section will present the method, the third section will show and discuss the results of the experiments, and lastly, the fourth section will conclude the research and future work.

2. The Proposed Method: Hybrid Naïve Bayes

HNB is a combination between lexical approach and Naive Bayes algorithm is one of the machine learning algorithm. Similar study by Alkubaisi et.al (2018) focuses on Twitter data, where as in our work, news data is used. The proposed HNB algorithm will follow 5 process steps namely: Data Collection, Data Preprocessing, Generating Labels for Training Dataset, Training the classifier and Classification. The steps are illustrated in Figure 1.

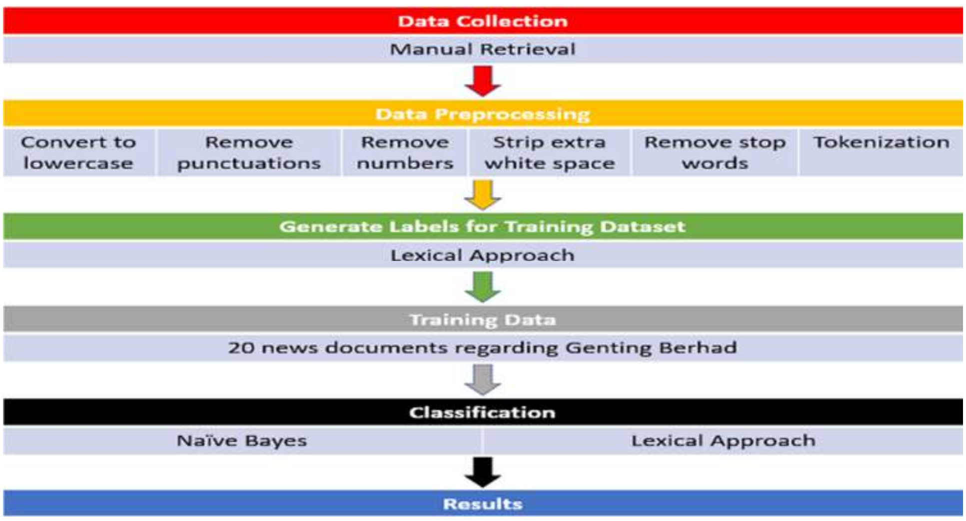

Figure 1. Process steps for hybrid Naïve Bayes

2.1. Data Collection
This research uses two types of data: the financial news and the historical stock price for the chosen company i.e. Genting Berhad, due to the availability of the resources needed for this study. The data was collected from two secondary sources as follows:
- The financial news was collected from *klse.i3investor.com*, which is a consolidated Malaysian stock market website, where it contains a forum, market blog site, live quotes, and financial news from all sources regarding a company.
- The historical stock price data was collected from *investing.com*.

The data was collected from the two sources over a period of 11 months, from 2nd January 2018 until 30th November 2018. Within this period, there were 223 days of stock price data from Bursa Malaysia and 41 articles released for Genting Berhad.

2.2. Data Pre-processing
Data pre-processing is an important step before applying the sentiment analysis. The collected news data is in unstructured free form; hence, the following data pre-processing steps was carried out:
- Text Normalization: The process of text normalization is referred to the removal of non-necessary characters. This process is carried out by converting the news into corpus, then converting the words in the corpus into lower case and lastly, removing the stop words, extra white spaces, numbers and punctuation marks.
- Tokenization: In the tokenization step, the string of the news is segregated into tokens based on the white space separator.

2.3. Generating Labels for Training Data
HNB is a supervised machine learning algorithm, which means labelled data is needed to train the algorithm. Lexical approach is use in labelling data for training. In lexical approach, the positive and negative words dictionary by [17] are used. The process applied is as follows:
1. Check whether the text document contains any positive or negative word in the dictionaries.
2. Assign "-1" score to each negative word and "+1" score to each positive word found in the text document.
3. Sum up the sentiment scores for the text document and the final sentiment score will indicate the news orientation (positive or negative).

3. Experiments
This section will elaborate the testing method, the results validation method and the results for the lexical approach and the NB algorithm.

3.1. Lexical Approach
The results of the lexical approach was used to check whether there is a relationship between the news on the internet and the stock price. In order to check the existence of the relationship, a correlation of the sentiment score extracted from the news and the historical stock price was calculated. To calculate the correlation, a normalization is needed. The sentiment score and the historical price were normalized using the min-max normalization $\bar{\boldsymbol{x}}$ which is illustrated in equation 1.

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{x} - \boldsymbol{x}\mathbf{min}}{\boldsymbol{x}_{\mathbf{max}} - \boldsymbol{x}\mathbf{min}} \quad \textbf{(1)}$$

The correlation coefficient r was calculated using the following equation:

$$\boldsymbol{r} = \frac{\boldsymbol{cov(x, Y)}}{\boldsymbol{\sigma_x \sigma_y}} \quad \textbf{(2)}$$

r gives a result fall in the period [-1,1] with -1 means perfect negative correlation, 0 means no correlation and 1 means perfect positive correlation.

3.2. NB Classifier

The dataset was separated into training and test sets, 60% for training and 40% for testing. The NB classifier was trained using the training dataset and tested using the test dataset.

3.3. Results and Discussion

The results of the correlation of the normalized stock price and news sentiment for Genting Berhad for the period of 2nd January 2018 through 30th November 2018 is shown in Figure 2. As can be seen from the Figure 2, there is a positive correlation between the news sentiment and the stock price. This means the news sentiment will affect the market's mood and it will also affect the stock price movement. The resulted correlation coefficient was 0.5841, indicating a moderate positive relationship.

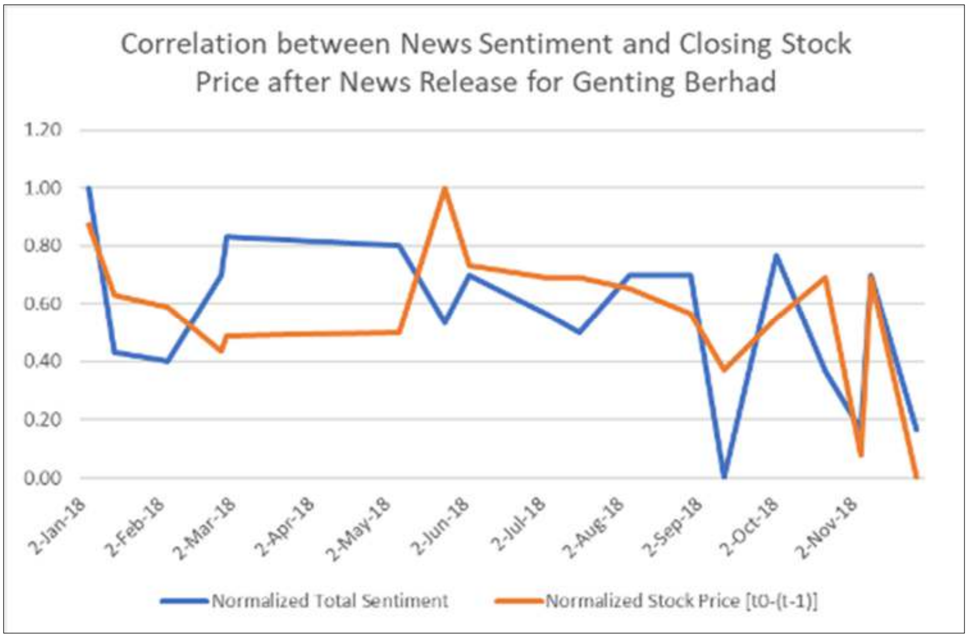

Figure 2. Correlation between News Sentiment and Closing Stock Price after News Release for Genting Berhad

For NB, the accuracy metric was used for results validation. The accuracy was formulated using the confusion matrix method. NB classifier give accuracy of about 69%. This shows that NB can be adopted as one of the machine learning algorithms in classifying news contents.

NB was used due to its fast performance, efficiency and that it requires a small training set, yet there are many of machine learning algorithms that can be used like Support Vector Machine (SVM) and Deep Learning techniques. These techniques are known for their high accuracy but also have drawbacks. The biggest drawback of SVM is model selection. For SVM, a kernel function have to be chosen and its parameter should be defined in advance [15] while deep learning needs a large number of training data and the training process is time consuming [16].

4. Conclusion and Future Work

This paper presented a research where Hybrid Naive Bayes was used to analyze a sentiment for financial news of Genting Berhad, from 2nd January 2018 to 30th November 2018. In answering RQ1, this study reveals that there exists a correlation of about 58.4% between news sentiment and historical stock price movement. This indicate a semi-strong relationship between the tone of the news and how it driven the stock price. In addition, Naive Bayes classifier gave a 69% of accuracy in classifying the news sentiments into polarity. The results indicated that Malaysian traders could trade their stocks, based on the sentiment data from the news released, as it can be considered as a reliable source of information, apart from the trend analysis, which answer RQ2.

This study was limited to news data of Genting Berhad. We hope to expand the study to cover the news for all the companies in the Malaysia stock market, and to use other social media sources such as Twitter and forum, in order to gauge the overall market mood. We plan to experiment with other machine learning techniques like deep learning, and propose attributes that will further improve its efficiency, apart from accuracy in performing the sentiment analysis.

References

[1] C. W. Leung, S. C. Chan, and F.-l. Chung, "Integrating collaborative filtering and sentiment analysis: A rating inference approach," in *Proceedings of the ECAI 2006 workshop on recommender systems*, 2006, pp. 62-66.

[2] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri, "Application of Deep Learning to Sentiment Analysis for recommender system on cloud," in *Computer, Information and Telecommunication Systems (CITS), 2017 International Conference on*, 2017, pp. 93-97: IEEE.

[3] F. Ali, E.-S. Shaker, A. Ali, K. Kwak, and D. J. 한. 학. Kwak, "Sentiment analysis of transportation using word embedding and LDA approaches," pp. 1111-1112, 2018.

[4] G. T. Giancristofaro and A. Panangadan, "Predicting sentiment toward transportation in social media using visual and textual features," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, 2016, pp. 2113-2118: IEEE.

[5] P. Sharma and T.-S. Moh, "Prediction of indian election using sentiment analysis on hindi twitter," in *Big Data (Big Data), 2016 IEEE International Conference on*, 2016, pp. 1966-1971: IEEE.

[6] C. S. Robertson, F. A. Rabhi, and M. Peat, "A service-oriented approach towards real time financial news analysis," in *Consumer Information Systems and Relationship Management: Design, Implementation, and Use*: IGI Global, 2013, pp. 32-49.

[7] J.-L. Seng and H.-F. J. K. Yang, "The association between stock price volatility and financial news–a sentiment analysis approach," vol. 46, no. 8, pp. 1341-1365, 2017.

[8] J. Kleinnijenhuis, F. Schultz, D. Oegema, and W. J. J. Van Atteveldt, "Financial news and market panics in the age of high-frequency sentiment trading algorithms," vol. 14, no. 2, pp. 271-291, 2013.

[9] J. Bollen, H. Mao, and X. J. J. o. c. s. Zeng, "Twitter mood predicts the stock market," vol. 2, no. 1, pp. 1-8, 2011.

[10] S. G. Chowdhury, S. Routh, S. J. I. J. o. C. S. Chakrabarti, and I. Technologies, "News analytics and sentiment analysis to predict stock price trends," vol. 5, no. 3, pp. 3595-3604, 2014.

[11] S. M. Shuhidan, S. R. Hamidi, S. Kazemian, S. M. Shuhidan, and M. A. Ismail, "Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm," in *International Conference on Kansei Engineering & Emotion Research*, 2018, pp. 64-72: Springer.

[12] G. A. A. J. Alkubaisi, S. S. Kamaruddin, H. J. C. Husni, and I. Science, "Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers," vol. 11, no. 1, p. 52, 2018.

[13] K. S. Komariah, C. Machbub, A. S. Prihatmanto, and B.-K. J. 멀. Sin, "A Study on Efficient Market Hypothesis to Predict Exchange Rate Trends Using Sentiment Analysis of Twitter Data," vol. 19, no. 7, pp. 1107-1115, 2016.

[14] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. J. a. p. a. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," 2016.

[15] G. T. Kaya, O. K. Ersoy, M. E. J. I. T. G. Kamasak, and R. Sensing, "Support Vector Selection and Adaptation for Remote Sensing Classification," vol. 49, no. 6-1, pp. 2071-2079, 2011.

[16] A. Voulodimos, N. Doulamis, A. Doulamis, E. J. C. i. Protopapadakis, and neuroscience, "Deep learning for computer vision: A brief review," vol. 2018, 2018.

[17] B. Liu, and M. Hu, Mining and Summarizing Customer Reviews. Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining, pp. 168-177, 2004