International Journal of Engineering & Technology, 7 (2.14) (2018) 57-61

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers

Ghaith Abdulsattar A. Jabbar Alkubaisi[1]*, Siti Sakira Kamaruddin[2], Husniza Husni[3]

School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Sintok 06010, Malaysia
**Corresponding author E-mail: ghaith.alkubaisi@outlook.com*

Abstract

Sentiment analysis has gained a lot of importance in last decade especially on the availability of data from Twitter that has created more interest for research in this field. Nevertheless, stock market classification models still suffer less accuracy and this has affected negatively the stock market indicators. In this paper, a new framework related to sentiment analysis from Twitter posts is proposed. The proposed framework represents an improved design of classification model that works to improve the classification accuracy to support decision makers in the domain of stock market exchange. This model starts with data collection part and in second phase filtration is done on data to get only the relevant data. The most important phase is the labelling part in which polarity of data is determined and negative, positive or neutral values are assigned to statements of people. The fourth part is the classification phase in which suitable patterns of stock market will be identified by hybridizing NBCs. The last phase is performance and evaluation. This study proposes to a Hybrid Naïve Bayes Classifiers (HNBCs) as a machine learning method for stock market classification, hence represents a useful study for investors, companies and researchers and will help them to formulate their policies according to sentiments of people.

Keywords: *Classification Accuracy; Naïve Bayes Classifiers; Sentiment Analysis; Stock Market Classification Model; Twitter*

1. Introduction

The selection of stock market models have always been a point of debate among researchers. Each model has its own strengths and weaknesses. Current stock market classification models are still suffering low accuracy in classification [41, 19, 7]. The low accuracy in classification have direct effects on the reality and reliability of stock market indicators like a series of statistical figures and financial reports which explain the stock behavior in existing stock market [6, 11, 27]. There are many factors that affect the accuracy of classification model results such as features of data, sample size, period of collecting data, and data classification techniques [5, 23, 32, 40]. Sentiment analysis helps a firm to analyze its customers' opinions and feedbacks.

Using this analysis, they can improve their services or product quality in the future that will not only help them to attract new customers but they will also retain current ones as well.

Tweets help users and organizations to collect valuable information in different domains8. In the recent years, one of the most popular research area using sentiment analysis on Twitter is the stock market [39].

This Study proposes two important features namely: temporal and spatial. Additionally, expert labelling will be employed as another step of the sentiment analysis to generate more suitable labelled data that will improve the output of classification. This is due to the fact that each stock market has specific feedback. Irrelevant words will be dropped and more concentration will be given on the goal of classification model to increase exactness of inputs data and build more related data before classification [4]. Finally, concerning the challenges of neutral ratio, this study will rely on baseline strategy that would be learned from the original training dataset to classify the polarity classes with a binary strategy that uses polarity lexicon to identify positive, negative and neutral tweets [17].

2. Related works

2.1. Twitter data source

[41, 29] have focused on the period of collecting dataset from Twitter to perform sentiment analysis. After applying the approach, it was observed that the changes in stock prices are correlated with individual words, then a stronger correlation to certain stocks can be built based on a dataset from those words. Expanding the period of collecting data and the size of keywords set without specifying the features of data are not enough to achieve the required accuracy and reliability for classification. Therefore, in this study two substantial features will be selected to achieve more accurate results for classification. These two substantial features are temporal and spatial features: - spatial feature to represents the Tweet's geographical location and temporal feature to represents the Tweet's timestamp [34].

2.2. Labelling technique

Automatic labelling is automatically identifying sentiment expressed in a given Tweet based on the prior general lexicon not specifically relevant to the research area [6, 21]. The research area includes stock market performance [12], crime prediction [36], and tourism information [33] to name a few. According to [3], the automatic labelling is still unable to tell the real public moods since the

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

data is usually collected directly from the logs. [9] Suggested that automatic labelling step could be improved for better sentiment calculations to achieve the required dataset which enhances the accuracy of the classification.

2.3. Polarity ratios

The main challenge in existing machine learning methods that are used in stock market classification model is the problem of neutral ratio. According to [31, 13] the conflicting ratio between neutral and positive classes was performed automatically by giving the positive polarity for the majority of the neutral set. Neutral rate affects the process of modelling and its main goal is to increase the reality and the accuracy of the pattern between public opinion on Twitter and the situation of stock in the stock market for the specific location [20]. Identifying the ratio of neutral polarity by the specialists in the field of the stock market during the labelling step is considered very important because it will affect the reality of classification [24, 5].

2.4. Classifiers

Nowadays, the supervised machine learning methods like NBCs are widely used to analyse the consumer reactions (user's opinion) from social networks [6, 16]. Naïve Bayes Classifiers (NBCs) are probabilistic supervised machine learning classifiers that use Bayes rule; all features included in the data are assumed features independence. NBCs have four models, Gaussian NB, Multinomial NB, Bernoulli NB and Semi-Supervised NB. Hybridizing algorithms of these classifier models with different numbers of variables and features lead to achieving the optimizations [30, 2].

3. The proposed method

A conceptual framework is developed to break down the implementation processes of generating an enhanced model for HNBCs. Figure 1 consists of five phases; firstly, the dataset from Twitter will be collected by using Twitter Application Program Interface (API) because the streaming can provide a continuous stream of the information with updates [18]. Secondly, pre-processing the data set using Natural Language Pre-processing (NLP) [1], where NLP processing start with transformations step [25] and end with the extraction process to extract the required features [40]. Thirdly, is the expert labelling technique where in this phase the dataset will be classified into positive, negative and neutral polarity. Fourthly, is the classification level using HNBCs is a baseline NB with combining Multinomial NB (MNB) as handling more than one feature at the same time [35], Bernoulli NB (BNB), which is very easy for more training data set [15] and Semi-Supervised Naïve Bayes (SSNB), which is suitable for train NBCs for labelled data [10]. Finally, the classification model's performance will be evaluated using recall, precision, F-measure and accuracy.

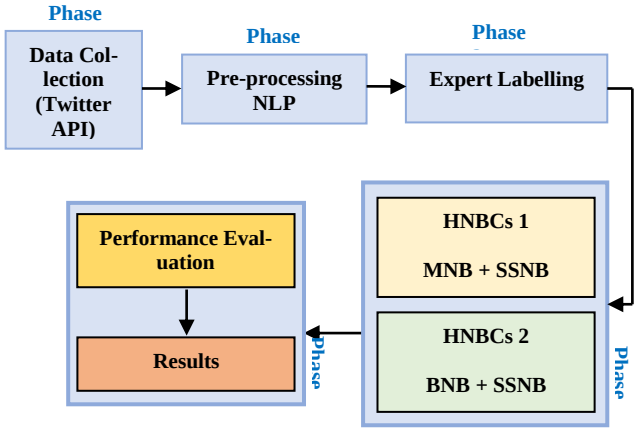


Fig. 1: A conceptual framework

Phase 1: Data collection (Twitter API)

This study will focus on tweets that reflect the consumer reactions about companies that provide services and products. In Twitter, there are two types of API's used to gather tweets, Twitter REST API and Twitter Streaming API [18]. This study will use an internal library using REST API, and will use Application Only Authentication (OAuth) required by Twitter4j library. OAuth is used in Twitter4j library Twitter and it supports access and provides authorized access to its API [38]. APIs streaming can provide a continuous stream of the information with updates. At the same time, the study has search API and streaming API for Twitter sentiment analysis. Streaming API can access real-time data of tweets using queries. More so, this study will choose search API which is a REST API since it enables users to retrieve recently posted tweets by specific queries using HTTP methods [28]. In addition, it can filter results based on time, regions, and languages26. The return of queries is a list of JSON objects containing tweets and metadata. These objects involve username, location, time and re-tweets [6].

Phase 2: Data Preprocessing

This study will adopt the whole steps of the pre-processing method starting from cleaning text, removing white space, extracting abbreviation, removing stop words, handling negation, and selecting feature. More so, the last process in pre-processing method works on extract all relevant contents from tweets and drop the irrelevant contents, for this reason, it is called a filtering step and all steps before it are called transformations [1, 6].

a. Transformations step includes the following steps [25]: (1) Tokenizing text by splitting it using spaces; (2) Removing stop words like or, also, etc; (3) Eliminating URLs, usernames, hashtags, Twitter symbols, punctuations, and references; and (4) reducing redundant letters such as "cooool" to "cool".

b. Filtering step is related to the content extraction process from the collected tweets after transformations step [40]. This study will focus on extracting two important features, which are spatial and temporal features of tweets. Spatial information about tweets can be obtained by using two ways: the first one is automatically collecting the accurate spatial information available on Twitter and the second one is approximately inferring the location of the user from the user profile [14, 22]. This study will also use shape let temporal selection. This type of feature selection assume that the time series are independent and identically distributed [37].

Phase 3: Expert labelling

This study advocates the expert labelling technique to define the polarity (positive, negative and neutral) of data after pre-processing phase. The labelling technique performed by experts with experiences in the domain of stock market plays the main role in enhancing the polarity result which leads to increasing the classification accuracy [23].

Phase 4: Classification

This study will specifically apply classification method to identify a suitable pattern in the domain of stock market classification model by hybridizing NBCs. NBCs have four parameter estimators that are Gaussian Naïve Bayes (GNB), MNB, BNB and SSNB. The hybridization of NBCs for this study depends on combining the following three algorithms MNB, BNB, and SSNB. This study chooses the three classifiers because their characteristics are matched to the requirements of it, for instance MNB handling more than one feature at the same time as this study is focusing on different types of features such as spatial and temporal. The pseudo code of HNBCs1 and HNBCs2 algorithm are presented as follows: -

- **HNBCs 1 (MNB+SSNB):**

Create a frequency (or tf-idf) table for all the features in the train set against every single classes every single document. Extra class refers to unknown data. Supposed known documents feature matrix is shown as

$$F(i,j,k)\ \ i = 1..\#class, j = 1..\#feature\ , k = 1..\#document_k \quad (1)$$

And unknown document feature matrix is shown as

$$UF(k,j)\ \ k = 1..\#document_{uk}, j = 1..\#feature \quad (2)$$

Determine the initial weight of every single known and unknown document.

$$W(i,k) = \begin{cases} 1, & \text{if } document_k \in class_i\} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Calculate the probability of the features against the classes.

$$\theta(i,j) = \frac{\sum_{k=1}^{\#document_i} W(i,k)F(i,j,k) + \alpha}{\sum_{k=1}^{\#document_i}\sum_{j=1}^{\#feature} W(i,k)F(i,j,k) + \alpha n},$$
$$i = 1..\#class, j = 1..\#feature \quad (4)$$

In special case $\alpha = 1$ and $n = \#feature$

Calculate each class probability.

$$P_c(i) = \frac{\#document_i}{\#all\ document},\ \ i = 1..\#class \quad (5)$$

Calculate all unknown document class probabilities.

$$P_k(i) = \log(P_c(i)).\sum_{j=1}^{\#feature} UK(k,j).\log(\theta(i,j)) \quad (6)$$

Calculate unknown data weights.

$$W(i,k) = \frac{P_k(i)}{\sum_{j=1}^{\#class} P_k(j)}\ \text{where } i = 1..\#class, k = 1..\#document_{uk} \quad (7)$$

Return equation 4 while the maximum iteration complete or there is no significant difference on unknown data weight.

Create a frequency (or tf-idf) table for all the features for certain test element. Suppose these feature matrix is shown as

$$F_t(l)\ \ l = 1..\#feature\ of\ test\ element \quad (8)$$

Calculate the conditional probabilities for all the classes, i.e..,

$$P(i) = \log(P_c(i)).\sum_{l=1}^{\#feature} F_t(l).\log(\theta(i,l)) \quad (9)$$

Determine the result which has maximum probability

$$result = \underset{i}{\operatorname{argmax}}(P(i)) \quad (10)$$

- **HNBCs 2 (BNB+SSNB):**

Create a binary table for all the features in the train set against every single classes every single document. Suppose these feature matrix is shown as

$$F(i,j,k)\ \text{where } i = 1..\#class, j = 1..\#feature\ , k = 1..\#document_k \quad (11)$$

And unknown document feature matrix is shown as

$$UF(k,j)\ \text{where } k = 1..\#document_{uk}, j = 1..\#feature \quad (12)$$

Determine the initial weight of every single known and unknown document.

$$W(i,k) = \begin{cases} 1, & \text{if } document_k \in class_i\} \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Calculate the probability of the features against the classes.

$$\theta(i,j) = \frac{\sum_{k=1}^{\#document_i} W(i,k)F(i,j,k) + \alpha}{(1+\alpha)\sum_{k=1}^{\#document_i} W(i,k)},$$
$$i = 1..\#class, j = 1..\#feature \quad (14)$$

In special case $\alpha = 1$ and $n = \#feature$

Calculate each class probability.

$$P_c(i) = \frac{\#document_i}{\#all\ document},\ \ i = 1..\#class \quad (15)$$

Calculate all unknown document class probabilities.

$$P_k(i) = P_c(i).\prod_{l=1}^{\#feature}\theta(i,l)UK(k,j) + (1-\theta(i,l))(1-UK(k,j)) \quad (16)$$

Calculate unknown data weights value.

$$W(i,k) = \frac{P_k(i)}{\sum_{j=1}^{\#class} P_k(j)}\ \text{where } i = 1..\#class, j = 1..\#document_{uk} \quad (17)$$

Return equation 4 while the maximum iteration complete or there is no significant difference on unknown data weight.

Create a binary feature vector for certain test element. Suppose these feature matrix is shown as

$$F_t(l)\ \ l = 1..\#feature\ of\ test\ element \quad (18)$$

Calculate the conditional probabilities for all the classes, i.e..,

$$P(i) = P_c(i).\prod_{l=1}^{\#feature}\theta(i,l)F_t(i) + (1-\theta(i,l))(1-F_t(i)) \quad (19)$$

Determine the result which has maximum probability

$$result = \underset{i}{\operatorname{argmax}}(P(i)) \quad (20)$$

Phase 5: Performance and evaluation

Measurements that would be used in this evaluation process to compare the proposed method with the latest studies that applied sentiment analysis based on NB classifiers.
Our equation is given below as:

- **a.** Recall = Sensitivity = Total Positive Rate is a proportion of cases that were correctly identified as positive. It is defined as [TP / (TP + FN)] = [d / (c + d)].
- **b.** Precision [TP / (TP+ FP)].
- **c.** False Positive Rate are the proportions of cases that were incorrectly identified or classified as positive. [b / (a + b)] or [FP/ (TP +FP)].
- **d.** Accuracy is defined as the portion or part of the sum total number of classification that is correct. It is given as [(a + d)/(a + b + c + d)] or [(TP + TN) / (TP + FP + FN + TN)].

4. Conclusion

The proposed model presented in this study goes through five phases. It starts with collection of data and then filtering the irrelevant tweets, after that the most important phase starts in which polarity is determined and negative, positive or neutral values are

assignments according to sentiments of people expressed in their tweets. The fourth step works to enhance the classification accuracy to support or to serve the decision makers in the domain of stock market exchange, according to sentiments of people by enhancing NBCs. Finally, is the performance and evaluation phase. Our proposed method is based on the use of sentiment analysis approach, by employing the expert labelling technique and spatial and temporal features. This study proposes a HNBCs as a machine learning method for stock market classification. A hybrid algorithm will be adapted from NB, where the hybrid algorithm will incorporate two different NB algorithms based on their specified functionalities.

References

[1] Abdelwahab, O., Bahgat, M., Lowrance, C. J., & Elmaghraby, A. (2015, 7-10 Dec. 2015). Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis. Paper presented at the 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).

[2] Aggarwal, C. C., & Zhai, C. (2012). Mining text data: Springer Science & Business Media.

[3] Ahuja, R., Rastogi, H., Choudhuri, A., & Garg, B. (2015). Stock market forecast using sentiment analysis. Paper presented at the Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on.

[4] Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of arabic tweets. International Journal of Social Network Mining, 2(2), 101-114.

[5] Alkubaisi, G. A. A., Kamaruddin, S. S., & Husni, H. (2017). A Systematic Review on the Relationship Between Stock Market Prediction Model Using Sentiment Analysis on Twitter Based on Machine Learning Method and Features Selection. Journal of Theoretical and Applied Information Technology, 95(24), 6924-6933.

[6] Alkubaisi, G. A. A., Kamaruddin, S. S., & Husni, H. (2018). Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. Computer and Information Science, 11(1), 52.

[7] Alm, E. C. O. (2008). Affect in text and speech: ProQuest.

[8] Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

[9] Arvanitis, K., & Bassiliades, N. (2017). Real-Time Investors' Sentiment Analysis from Newspaper Articles Advances in Combining Intelligent Methods (pp. 1-23): Springer.

[10] Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. Computational Intelligence, 31(1), 132-164.

[11] Attigeri, G. V., MM, M. P., Pai, R. M., & Nayak, A. (2015). Stock market prediction: A big data approach. Paper presented at the TENCON 2015-2015 IEEE Region 10 Conference.

[12] Bhattu, N., & Somayajulu, D. (2012). Semi-supervised Learning of Naive Bayes Classifier with feature constraints. Paper presented at the 24th International Conference on Computational Linguistics.

[13] Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. Computer, 44(10), 91-94. doi: 10.1109/MC.2011.323

[14] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[15] Cakra, Y. E., & Trisedya, B. D. (2015). Stock price prediction using linear regression based on sentiment analysis. Paper presented at the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

[16] Chandra, S., Khan, L., & Muhaya, F. B. (2011). Estimating Twitter user location using social interactions--a content based approach. Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.

[17] Di Nunzio, G. M., & Sordoni, A. (2012). A visual tool for bayesian data analysis: the impact of smoothing on naive bayes text classifiers. Paper presented at the Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.

[18] Fiarni, C., Maharani, H., & Pratama, R. (2016, 25-27 May 2016). Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. Paper presented at the 2016 4th International Conference on Information and Communication Technology (ICoICT).

[19] Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2013). TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. Paper presented at the Workshop on Sentiment Analysis at SEPLN (TASS2013).

[20] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, 12.

[21] Hajli, M. N. (2014). The role of social support on relationship quality and social commerce. Technological Forecasting and Social Change, 87, 17-27.

[22] Hamed, A.-R., Qiu, R., & Li, D. (2015). Analysis of the relationship between Saudi Twitter posts and the Saudi stock market. Paper presented at the 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS).

[23] He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. Information Processing & Management, 47(4), 606-616.

[24] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012). Discovering geographical topics in the Twitter stream. Paper presented at the Proceedings of the 21st international conference on World Wide Web.

[25] Jiang, L., Wang, D., Cai, Z., & Yan, X. (2007). Survey of improving naive Bayes for classification. Paper presented at the International Conference on Advanced Data Mining and Applications.

[26] Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. Computational Intelligence, 22(2), 100-109.

[27] Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! ICWSM, 11(538-541), 164.

[28] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C.-C. (2012). Tedas: A Twitter-based event detection and analysis system. Paper presented at the 2012 IEEE 28th International Conference on Data Engineering.

[29] Lin, J., & Ryaboy, D. (2013). Scaling big data mining infrastructure: the Twitter experience. ACM SIGKDD Explorations Newsletter, 14(2), 6-19.

[30] Makice, K. (2009). Twitter API: Up and running: Learn how to build applications with the Twitter API: " O'Reilly Media, Inc.".

[31] Makrehchi, M., Shah, S., & Liao, W. (2013). Stock prediction using event-based sentiment analysis. Paper presented at the Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on.

[32] Nguyen, T. T. T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials, 10(4), 56-76. doi: 10.1109/SURV.2008.080406

[33] Qasem, M., Thulasiram, R., & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets. Paper presented at the Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on.

[34] Sathyadevan, S., Sarath, P. R., Athira, U., & Anjana, V. (2014, 26-28 Aug. 2014). Improved document classification through enhanced Naive Bayes algorithm. Paper presented at the Data Science & Engineering (ICDSE), 2014 International Conference on.

[35] Shimada, K., Inoue, S., Maeda, H., & Endo, T. (2011). Analyzing tourism information on Twitter for a local city. Paper presented at the Software and Network Engineering (SSNE), 2011 First ACIS International Symposium on.

[36] Song, Z., & Xia, J. C. (2016). Spatial and Temporal Sentiment Analysis of Twitter data. European Handbook of Crowdsourced Geographic Information, 205.

[37] Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34(4), 2622-2629.

[38] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. Paper presented at the Proceedings of the ACL 2012 System Demonstrations.

[39] Wang, H., Wu, J., Zhang, P., & Zhang, C. (2016). Temporal Feature Selection on Networked Time Series. arXiv preprint arXiv:1612.06856.

[40] Yamamoto, Y. (2014). Twitter4J-A java library for the Twitter API: sep.

[41] Yan, D., Zhou, G., Zhao, X., Tian, Y., & Yang, F. (2016). Predicting stock using microblog moods. China Communications, 13(8), 244-257. doi: 10.1109/CC.2016.7563727

[42] Yang, A., Zhang, J., Pan, L., & Xiang, Y. (2015, 16-18 Nov. 2015). Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination. Paper presented at the Security and Privacy in Social Networks and Big Data (SocialSec), 2015 International Symposium on.
[43] Zhang, L. (2013). Sentiment analysis on Twitter with stock price and significant keyword correlation (Doctoral dissertation).