DOI: 10.7763/IPEDR. 2012. V55. 8

Stock Price Predication using Combinational Features from Sentimental Analysis of Stock News and Technical Analysis of Trading Information

Jheng-Long Wu [1], Chen-Chi Su [1], Liang-Chih Yu [1 +] and Pei-Chann Chang [1]

[1] Department of Information Management, Yuan Ze University, 135 Yuan-Tung Rd., Taoyuan, 32026, Taiwan

Abstract. Stock trend prediction is an appealing research problem to be investigated. Many researches use technical analysis (TA) to predict stock trend but it is very easy affected i.e. stock message and policy. Stock news articles play an important role in stock trend prediction because these articles affect investors' decision on stock investment. Text mining techniques are useful to find valuable information from textual corpora, but many researches use dictionary by expert definition only. In this paper, we propose a stock price predication model which is combinational feature from technical analysis and sentiment analysis (SA). The features of sentiment analysis is based on a Pointwise mutual information (PMI) which is a term expansion method from multidimensional seed word. The features of technical analysis based on expert rule from trading information. Experimental results show that the use of sentiment analysis and technical analysis achieves higher performance than that without sentiment analysis in predicting stock price.

Keywords: Pointwise Mutual Information, Sentiment Analysis, Technical Analysis, Stock Price Predication

1. Introduction

Stock information has multiple categories, i.e. stock closed price, trading volume, stock news, stock message, and expert analysis. Many studies have been done to predict the stock price by using statistics and machine learning techniques using historical stock price and trading volume [1]. In general, technical analysis (TA) is based on historical developed regularities in the stock exchange with an assumption that the same result will repeat in the future [2-3]. There are many influential indicators and trading rules based on them. Technical indicators might provide advice to traders on whether a trend will continue, such as MACD, or whether a stock is oversold or overbought, such as BIAS. One of the important issues for forecasting market trend is to know sentiment of stock news, that it's good or bad trend, when the financial stock prices go through the up/down cycle. The sentimental analysis (SA) can be applied to make trading decisions where some of potential important information affecting investor to investment. The news sentimental analysis can use text mining technique to find best information [4-7]. SA has been developed and good performance in many researches [8-10]. The sentimental analysis is a different way to mining stock information compares to uses the trading information to predict future stock trends. In addition, many researchers using textual information to improved prediction performance [11-13]. However, these approaches have a problem that textual data is high complex information representation, whether using dictionary or manual lexicon by analyser may miss many of distinctive features. Therefore, the feature extraction can capture more effective variables as information to improved classification or predication problem [14-17], such as using SentiWordNet, association rule mining (ARM), Pointwise mutual information (PMI) and mutual information (MI).

To resolve the numerical prediction, we relied on knowledge learning and techniques from computational intelligence to reduce investment risks. There are many poplar tools to predict numerical value such as neural network (NN) and support vector regression (SVR). Presently, support vector regression, which was evolved from support vector machine (SVM) based on the statistical learning theory. This is a

[+] Corresponding author. Tel.: + 88694638800 ext. 2789; fax: +88634352077.
E-mail address: lcyu@saturn.yzu.edu.tw.

machine learning approach which has powerful forecasting, high learning capability and accuracy for numerical prediction [18-20].

In this paper, our research aim is to building a stock price predication system using combinational features from sentimental analysis of stock news and technical analysis of trading information In addition, we consider multidimensional sentimental intensity into sentimental analysis by expanding sentiment features based on multiple seed word sets. However, we use two analysis methods to generate affecting features for improving predicting performance. The major processes is to including calculating technical indices by technical analysis, mining effective feature have highly associated by PMI method and training a predication model based on two combinational feature sets from sentimental analysis and technical analysis..

2. A Stock Price Prediction Model based on Sentiment Analysis and Technical Analysis

The main objective is developing stock price predication model. From the Figure 1, there are four parts of this framework: 1) pre-processing the textual data for detecting the seed words, then mining the sentimental features by PMI measurement which has same meaning with seed and the features weighting as well as suing PMI measurement to determining, it can also generate feature set of sentiment analysis, 2) we calculate the technical indices based on price and volume form trading information on stock exchange for generate feature set, 3) using a machine learning method to learning predication model based on combinational feature sets, 4) finally we can predict daily future stock price by predication model. The detail processes as follows:

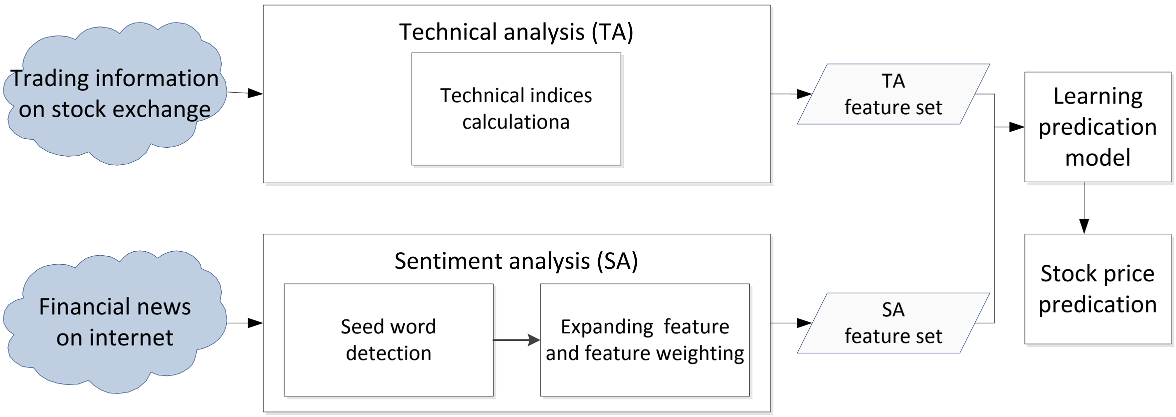
Trading information on stock exchange
Technical analysis (TA)
Technical indices calculationa
TA feature set
Learning predication model
Financial news on internet
Sentiment analysis (SA)
Seed word detection
Expanding feature and feature weighting
SA feature set
Stock price predication

Fig. 1: The stock price predication model based on sentiment analysis s with technical analysis.

2.1. Sentiment Analysis Based on Stock News

- Data pre-processing and detecting the seed word set

The textual data collect from online stock news and use part-of-speech method to tagging each word from CKIP system (word segmentation on Chinese word). We select important word according to POS tagging including verb, noun and adjective and then generate multidimensional seed word seta according multidimensional considerations i.e. economy, technology. Each seed word set is detecting by specific field expert because of the reason that human can identify sensitively the seed word according to their background in the field expert.

- Extracting sentiment features and its weighting by PIM-based

In features extraction, we want to use PMI method [9] to analyze word association among word and seed word set. Each word has PMI value from seed word set from last step. The value of *PMI (word, sword)* which is the *word* with the *sword* seed word calculates follows as:

$$PMI(word, sword) = \log_2\left(\frac{count(word, sword)}{count(word)count(sword)}\right) \tag{1}$$

where *count*(word, *sword)* is count the co-occur frequency between words.

Form the PMI value we calculate the strength of semantic association between *word* and seed word set of *Class* (i.e. positive, negative). The word strength is following as:

$$strength(word) = \frac{1}{n} \sum_{sword_i \in Class}^{n} PMI(word, sword_i) \tag{2}$$

where if the value in this *class$_1$* more than another *class$_2$* is belongs to the *class$_1$*, contrariwise belongs to *class$_2$*. Also the value of *strength (word)* is the feature weight of word in its *class*. Therefore, we could know the word that how many similarities with seed word set of class and each word has strength value. In addition, we can repeat the feature extraction processes from each seed word sets.

- Calculating sentiment intensity for each news

In this step, we provide a function to calculate sentimental intensity which is total stock information in a stock news document. The value of *Intensity* is a balance among positive and negative. The *p_strength* is positive feature detecting information volume of positive and the *n_strength* is negative feature detecting information volume of negative

$$Intensity(k) = p\_strength(k) - n\_strength(k) \tag{3}$$

$$p\_strength(k) = \frac{\sum_{i=1, i \in P, i \in D_k}^{n} strength(word_i)}{n} \tag{4}$$

$$n\_strength(k) = \frac{\sum_{j=1, j \in N, j \in D_k}^{m} strength(w_j)}{m} \tag{5}$$

where the *word$_i$* is the word strength of i^{th} feature in positive feature set *P* exist document *D$_k$*. *word$_j$* is the strength of j^{th} feature in negative feature set *N* exist document *D$_k$*. The *n* and *m* are total number of positive and negative feature appear in stock news *D$_k$*. According to the intensity of stock information determination the new affect degree by our proposed sentimental analysis, if the *Intensity* value more than zero then it has positive sentiment otherwise is negative sentiment.

2.2. Technical Analysis based on Stock Price and Volume

Investment managers calculate different indicators from available data and plot them as charts. Observations of price, direction, and volume on the charts assist managers in making decisions on their investment portfolios According to references that their researches shown the technical indices with stock price have high correlation coefficient. There are many kind of technical index in the stock market for investor decision which considers three major kinds of technical indices including moving average (MA), bias (BIAS) and relative strength index (RSI).

2.3. Learning Predication Model based on Technical Indices and Sentiment Intensity

In this section, we combine two feature set from sentiment analysis and technical analysis for prediction model. The sentimental analysis component can analyse the sentimental intensity of news in a day. The technical analysis component can generate technical indices in a day. The combinational feature set as input data is generated from SA and TA for learn the predication model. The target output is future stock price. Support vector regression will be applied as a machine learning model which can extract the hidden knowledge according to SA and TA. On the kernel function selection, we try to use RBF functions to generate better performance in SVR model..

2.4. Predicting the Daily Future Stock Price

In this part, we will calculate average sentimental intensity of stock news of each dimension of each day and combining the technical indices to stock price predication model. In this paper, we propose predict daily stock price based on our proposed predication mode.

3. Experimental Results

3.1. Datasets, Performance Evaluation and Experimental Design

The total textual data contains 7274 news in Chinese word which was collected from the website of Yahoo stock news. The data split to two data sets including trading data set has 5930 news from 06/01/2011 to 12/30/2011 and testing data set has 1344 news from 01/02/2012 to 04/20/2012. In Taiwan, many investors investment stock decision is well according to news form online news, therefore it forming a representative corpus to analyse the sentimental intensity of stock news. In addition, there are three feature set combinations for testing predication performance. One is SATA_all which combine two features of sentimental analysis and technical analysis; second is SATA_seed that combine technical analysis with seed word of sentimental analysis. Third is TA that using technical analysis only. In addition, we design incremental expansion feature sets for capture best feature expansion rate.

3.2. Predicting Results in Taiwan's Stock Market

In the experimental results of testing data in Table 1, the evaluation measures shows our feature extraction approach of sentimental analysis has lowest MAE which is 27.78 better than using both of technical analysis only and seed word of sentimental analysis only. As the results, our optimal expansion ratio are 10% from dimension 1 (Economy), 30% form dimension 2 (Science) and 50% from dimension 3 (Politics) but without investor dimension. More than dimension is investor. From the Fig. 2 shows that if expansion ratio last than 50% has stable performance and using 3 seed set of dimensions to expansion has bests result. Fig. 3 shows the price error on SATA_all and TA feature set which our combining feature set has least error.

Table 1: The stock price predication model based on sentiment analysis s with technical analysis.

Feature set	MAE	MAPE	RMSE
SATA_All	27.78	0.37%	35.64
SATA_seed	28.79	0.38%	35.84
TA	42.14	0.57%	51.51

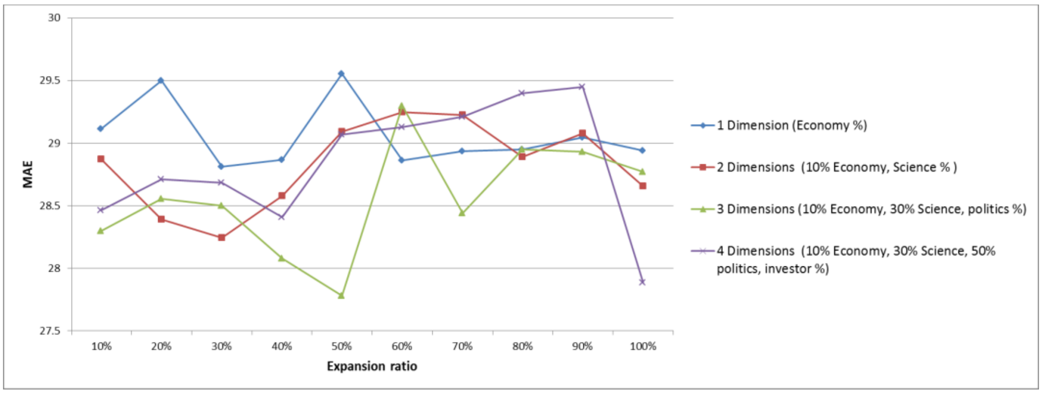

Fig. 2: The different expansion ratio on 4 dimensions based on SATA_All feature set.

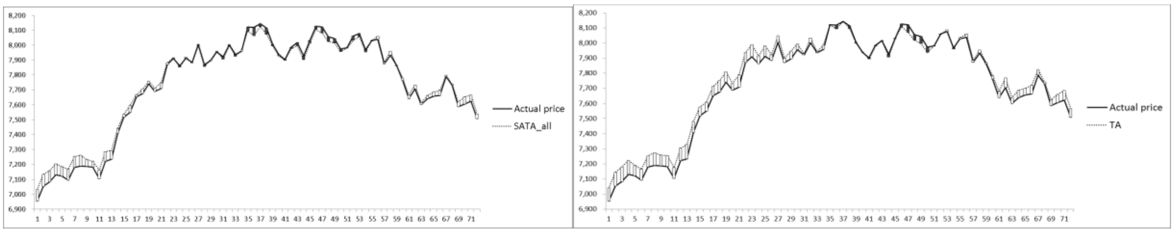

Fig. 3: Actual price compare to SATA_All and TA predication on Taiwan's TAIEX.

4. Conclusion

In this study, we design various feature set for stock price predication. Our proposed feature extraction approach of sentiment analysis can improve the predication performance. In addition, the experimental result shows our proposed model can improve the predicating performance that is PMI-based term expansion can capture effective sentiment feature from stock news. The multidimensional seed set can help feature expansion which has overall features. Therefore, the stock news can provide effective stock information for any predication model or combine other feature set. In the future work, we want to further analyse the relationship between seed word and expansion word i.e. sentence pattern, sequence and word distribution.

5. References

[1] P.-C. Chang and C.-H. Liu. A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications*. 2008, **34** (1): 135-144.

[2] M. A. H. Dempster, T. W. Payne, Y. Romahi and G. W. P. Thompson. Computational learning techniques for intraday FX trading using popular technical indicators. *IEEE Trans. on Neural Network*. 2001, **12** (4): 744-754.

[3] W. Lo, H. Mamaysky and J. Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance*. 2000, **55** (4): 1705-1765.

[4] J.-L. Wu, H.-S. Chu, L.-C. Yu and P.-C. Chang. Sentiment analysis of stock news using a PMI-based term expansion method. *ICIC Express Letters*. 2012, **6** (2): 491-496.

[5] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*. 2011, **50** (2): 491-500.

[6] S. S. Groth and J. Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*. 2011, **50** (4): 680-691.

[7] W. K. Chan and James Franklin. A text-based decision support system for financial sequence prediction. *Decision Support Systems*. 2011, **52** (1): 189-198.

[8] I. Maks and P. Vossen. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*. 2012, **53** (4): 680-688.

[9] N. Li and D. D. Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*. 2010. **8** (2): 354-368.

[10] X. Bai. Predicting consumer sentiments from online text. *Decision Support Systems*. 2011, **50** (4): 732-742.

[11] M. Cecchini, H. Aytug, G. J. Koehler and P. Pathak. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*. 2010, **50**(1): 164-175.

[12] R. P. Schumaker, Y. Zhang, C.-N. Huang and H. Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*. 2012, **53** (3): 458-464.

[13] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada and A. Sakurai. Combining technical analysis with sentiment analysis for stock price prediction. *Proc. of the 2011 IEEE International Conference on Dependable, Autonomic and Secure Computing*. 2011. pp. 800-807.

[14] J.-L. Wu, L.-C. Yu and P.-C. Chang. Emotion classification by removal of the overlap from incremental association language features. *Journal of the Chinese Institute of Engineers*. 2011, **34** (7): 947-955.

[15] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proc. of 7th Conference on Language Resources and Evaluation*. 2010. pp. 2200-2204, 2010.

[16] L.-C. Yu, C.-L. Chan, C.-C. Lin and I.-C. Lin. Mining association language patterns using a distributional semantic model for negative life event classification. *Journal of Biomedical Informatics*. 2011, **44** (4): 509-518.

[17] V. Vapnik. *The nature of statistical learning theory,* New York: Springer–Verlag, 1995.

[18] B. J. Chen, M. W. Chang, and C. J. Lin. A study on EUNITE competition 2001. *IEEE Trans. Power Syst*. 2004, **19**: 1821-1830.

[19] C.-J. Lu, T.-S. Lee and C.-C. Chiu. Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*. 2009, **48** (2): 115-125.

[20] H. Wang and A. S. Weigend. Data mining for financial decision making. *Decision Support Systems*. 2004, **37** (4): 457-460.