# Combining News and Technical Indicators in Daily Stock Price Trends Prediction

Yuzheng Zhai, Arthur Hsu, and Saman K Halgamuge

Dynamic System and Control Group,
Department of Mechanical and Manufacturing Engineering,
University of Melbourne, Victoria, Australia 3010
y.zhai@pgrad.unimelb.edu.au, {alhsu, saman}@unimelb.edu.au

**Abstract.** Stock market prediction has always been one of the hottest topics in research, as well as a great challenge due to its complex and volatile nature. However, most of the existing methods neglect the impact from mass media that will greatly affect the behavior of investors. In this paper we present a system that combines the information from both related news releases and technical indicators to enhance the predictability of the daily stock price trends. The performance shows that this system can achieve higher accuracy and return than a single source system.

## 1 Introduction

Stock market prediction has always been one of the hottest topics in research, as well as a great challenge due to its complex and volatile nature. Research suggests that the financial time series do not exhibit random behavior and the stock price is predictable [28]. Numerous publications have attempted to construct an accurate model for the stock market. Most of these works focus on time series prediction with various AI models, such as Artificial Neural Networks [1, 17], Genetic Algorithm [10], Fuzzy System [18], Hidden Markov Model [9] or some hybrid combinations [26, 29], as well as statistical techniques, such as moving average [6]. However, these methods inevitably have their own limitations. Back-propagation neural network for example, suffers from the risk of over-fitting and large number of parameters. More importantly, they have neglected other source of information such as mass media that will greatly affect the behavior of investors.

The entities listed on the Australian stock exchange are required to fully inform the investors at all times so that investment decision can be made with rich and timely information. The materials include negotiations of purchase, director appointment/resignation and divestitures of businesses. Since most of this information can be obtained from the news articles, major financial newspapers become a good source of information in assisting the traders. Mitchell and Mulherin [13] studied the influence of public information reported by Dow Jones and concluded that a direct relation does exists between released news articles and stock market activities. News release provides abundant information regarding the activities that companies are involved in and it may produce speculations among traders that results in movements of the stock prices. NofSinger [16] showed that in some cases, investors tend to buy

D. Liu et al. (Eds.): ISNN 2007, Part III, LNCS 4493, pp. 1087–1096, 2007.

after positive news which results in buying pressure and push the price higher; and sell after negative news which results in a drop in price. While there is no doubt that news releases create expectations among investors, there are only a few researches conducted recently in predicting the price movement using this information. Mittermayer [14] proposed a trading system to predict stock price trends immediately after the release of a news article through text mining techniques. They found the system significantly outperforms a random trader. However, only news that is directly related to the stock is included in their study while NofSinger's study suggests that both the firm specific and the general economics news affect trading behaviors [16].

However, using investors' expectations caused by news alone as a trading strategy is inadequate, as concluded by Brown and Cliff [23]. Therefore, this paper proposes to combine the information from both the news release and technical indicators to enhance the predictability of the daily stock price trends. The news articles used are composed of both company specific and relevant market sector news to reflect the overall impact of the media. Text mining techniques are employed to encode the news articles by forming and extracting important concepts. While support vector machine (SVM) [2], a supervised learning method, is applied as classifier. The resulting system is shown to be more accurate than the one that uses only single source of information.

The rest of the paper is organized as follow. Section 2 briefly introduces SVM and its application in financial and text mining. Section 3 presents the architecture of the system and design of various components. Research data and experiment results are detailed in Section 4. Conclusion and future research directions are given in Section 5.

## 2 Support Vector Machine

The SVM, originated form the work of Vapnik [22], is now being widely used to solve classification and prediction problems. SVM performs classification by constructing a hyperplane that separates the input space into two classes. It attempts to find the maximum margin hyperplane so that the separation between the decision classes is maximized. The input vectors that define the width of maximum margin are called support vectors and all the other points are not important in defining the separation. SVM maps the original input feature space into a higher dimension so that it can be separately by a linear model in the high dimensional space.

However, not all the problems can be separated by this *hard margin*. Therefore, in case of non-separable feature spaces, a *soft margin* is applied to allow some points to be misclassified. It chooses a hyperplane that separates the inputs as cleanly as possible, while still maximizing the distance between the support vectors. a penalty parameter of the error term $C$, ($C > 0$) can be set to be the upper bound in order to control the amount of deviation to be tolerated.

SVM has several advantages: it has little control parameters that need to be selected, over-fitting is unlikely to occur and it does not get trapped in a local optimum. It has been used in both areas of financial forecasting and text mining [8, 11]. There is a number of researches that has apply SVM to financial time series predictions and showed that SVM outperforms back-propagation networks [11, 20]. Dumais et al and Joachims demonstrated the applicability of SVM to text clustering applications and also suggest that SVM has outperformed others [3, 8].

## 3 System Architecture and Design

An overview of the system architecture is shown in Figure 1. The final prediction is determined base on two factors: the forecasting of price movement based on the technical indicators extracted from historical data and the combined impact of direct and indirect news articles related to the stock.

### 3.1 Price Forecast

Seven technical indicators [11, 24] are selected and computed for each trading day from the prices in the past five days. Table 1 presents the formula for each feature.

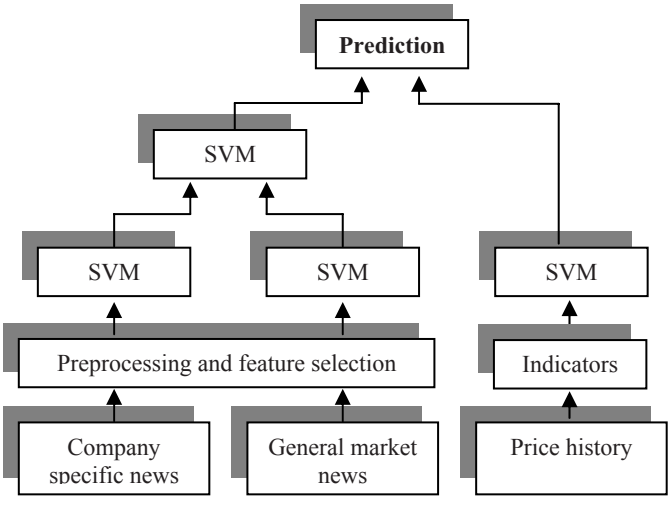


**Fig. 1.** Overview of the system architecture

**Table 1.** Summary of price features

| Feature | Formula | Feature | Formula |
|---|---|---|---|
| Stochastic %K | $\frac{C_t - LL_n}{HH_n - LL_n}$ | William's %R | $\frac{H_n - C_t}{H_n - L_n} \times 100$ |
| Stochastic %D | $\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$ | A/D Oscillator | $\frac{H_t - C_{t-1}}{H_t - L_t}$ |
| Momentum | $C_t - C_{t-4}$ | Disparity 5 | $\frac{C_t}{MA_5} \times 100$ |
| Rate of Change | $\frac{C_t}{C_{t-n}} \times 100$ | | |

$C_t$ is the closing price at day t, $L_t$ is the lowest price at day t, $H_t$ is the highest price at day t, $MA_n$ is the moving average of the past n days, $LL_n$ and $HH_n$ is the lowest low and highest high in the past n days, respectively.

The direction of next day's price movement is categorized into two classes: '1' or '-1'. '1' indicates the next day's closing price is higher or equal to today's closing price, and '-1' indicate a drop in share price.

### 3.2 News Prediction

There are two groups of news releases: the ones that directly relate to the stock and the ones that relate to the general market. Each group is trained separately as they may have different measure of influence on the stock. The categorization of the output classes is the same as in price forecasting. News is assumed to have valid influence on the stock only on the same day it is published.

The vector space model [19] is a commonly used technique in text mining, which has been successfully used in document categorization [12]. This model is employed in this study to represent the document in high dimensional space. It represents each document as binary vectors where each element is a word from a vocabulary. The elements will have a value of 1 if the corresponding word is present within the document or have a value of zero otherwise. A weight can be associated with each element to reflect their relative importance.

The preprocessing stage starts by first removing common stop words (such as "the", "a", etc) from each documents, and then the remaining words are tagged with their corresponding Part-Of-Speech (POS) tag. Instead of using each word directly, a background thesaurus WordNet [4] is used to replace words by higher level concepts [7]. WordNet is a semantic network that can give hierarchical hypernyms and hyponyms relations between words. The use of concepts increases the flexibility of the system to be able to account for vocabulary changes, as well as reduces the dimensions of feature space. The POS tag generated earlier is utilized to help disambiguating the word when assigned to WordNet.

The concepts are weighted by the conventional multiplicative combination of term frequency (TF) and the inverse document frequency (IDF), so that terms occur more often in a document and/or rarer in other documents will be given a higher weight. Moreover, those concepts that only occur in one class but not in the other are given a higher weight (multiplied by an arbitrary constant, 2 in this case) to help better distinguish between classes. Examples of some the unique concepts from documents that are considered as "good news" are: establishment, accumulation, growth, etc; while the ones from "bad news" are: separate, discharge, impair, etc. The feature space is then reduced to be the top 30 concepts with the highest weights, which are used to code each document.

The two groups of news are trained and classified using SVM and their results feed into another SVM to produce the combined prediction of price trends.

### 3.3 SVM

In this study, Gaussian RBF kernel and polynomial kernel are used for SVM. The only controlling parameters are the upper bound $C$ and the feature width $\sigma$ in case of BRF or the power $d$ in case of polynomial. These parameters are varied to ensure the optimal values are selected. Table 2 below presents an example of the performance of SVM with different parameters on the price data set.

**Table 2.** Prediction performance with various parameter values

| Parameter value | Hit ratio (%) |
|---|---|
| $C = 10$ | |
| $\sigma = 1$ | 35.5 |
| $\sigma = 3$ | 58.8 |
| $\sigma = 5$ | 50.0 |
| $\sigma = 7$ | 50.0 |
| $\sigma = 3$ | |
| $C = 1$ | 50.0 |
| $C = 20$ | 61.7 |
| $C = 50$ | 52.9 |
| $C = 10$ | |
| $d = 2$ | 57.6 |
| $d = 3$ | 52.9 |
| $d = 4$ | 51.9 |
| $d = 2$ | |
| $C = 1$ | 38.8 |
| $C = 20$ | 47.1 |
| $C = 50$ | 44.1 |

The results obtained conforms to Tay and Cao's findings where the prediction performance deteriorates when the value of $C$ and $\sigma$ are either too small or too large [20]. Finally, the value of $C$ is set to be 20, $\sigma$ to be 3 and $d$ to be 3.

Gaussian RBF kernel is used most of the time as it performs slightly better and takes less time. However when classifying the news articles, it often reaches 100% classification rate for one class, yet 0% for the other. Therefore, polynomial kernel is used instead in this case as it is less biased toward one class.

## 4 Experiment and Results

The research data used in this study is the daily prices (open, high, low, close) of BHP Billiton Ltd. (BHP.AX) of Australian Stock Exchange between March 1st, 2005 and May 31st, 2006. As well as the news articles related to BHP and its market sector in the same period that are published on Australian Financial Review, a major newspaper on business, finance and investment news in Australia. BHP is chosen because not only it has a large trading volume (20 Million on average) so it can be assumed that the transactions can take place whenever required, but also due to its popularity that attracts a lot of media attentions. Since BHP mainly involves in material mining, articles that concerns the directions of the metal market are included as general economic news. However, any press releases that only report the outcome of the day before are specifically excluded as it gives no new information. The data points in the first 12 months were used as training set, while the remaining two months serves as validation set. There are 286 training data and 34 holdout data from the price section. And for the news section, there are 120 training data and 28 holdout

data within the direct news category, while 53 training data and 15 holdout data are in the indirect news category. Table 3 below shows the prediction accuracy for using different data sets as inputs.

**Table 3.** Prediction accuracy

| Data sets | Accuracy (%) |
|---|---|
| Price | 58.8 |
| Direct news | 62.5 |
| Indirect news | 50.0 |
| Combined news | 64.7 |
| Price & News | 70.1 |

While the prediction accuracy is comparable to that obtained by Kim [11] using technical indicators only (on a different stock index), it is clear that with the combined information from both time series and textual data, the performance of stock trend prediction is noticeably improved.

It is also observed that when the predictions made from the numerical data are in conflict with the predictions from news articles, the later is more accurate most of the time. In the training data set, there are 39 inconsistent predictions, and the combined news predictions are correct for all of them and in the testing data, the accuracy is four out of six. Thus, news predictions in this case are considered more important than the data predictions.

### 4.1 Market Simulation

A market simulation is conducted to evaluate the profitability of the system under real life conditions. It is assumed that the initial investment is $10,000 and each transaction (buy/sell) incurs a fee of $20. The two months validation data (April and May 2006) used in previous section are employed in this test. Figure 2 below shows the price movements of BHP during those two months, which is a reasonably representative example. Further more, in order to avoid excessive transaction charges that will result from frequent operations, each day is limited to one transaction (buy/sell) only.

Three sets of similar strategies are used in this study for different input sets. If the prediction is based on past prices only, then the strategy for buy, sell and hold follows five simple rules:

- If the predicted trend is positive and the share has been bought, then hold.
- If the prediction is negative and the share hasn't been bought, then do nothing.
- If the prediction is positive and the share hasn't been bought, then buy.
- If the prediction is negative and the share has been bought, then sell.
- All transaction take place at the end of each trading day, thus the closing price is assumed to be the trading price.

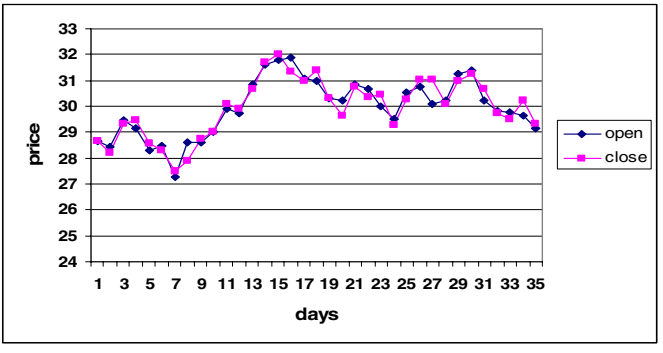


**Fig. 2.** Plot of closing/opening prices vs. date of BHP.ax during April and May, 2006

If the prediction is based on news only, then the above tactics are no longer applicable. This is due to the fact that news articles for each day can only be obtained in the morning before stock market opens. Therefore, changes are made to accommodate for the late arrival of news input and it is summarized as follows:

- Assuming the overall trend of the stock is rising (this assumption certainly may not be valid and it is solely used to simplify the experiment)
- If the news prediction is positive and the share hasn't been bought, then buy at opening price.
- If the prediction is positive and the share has been bought, then hold.
- If the prediction is negative and share hasn't been bought, buy at closing price.
- If the prediction is negative and the share has been bought, sale at opening price.
- If the news prediction is absent, then buy at closing price.

A recent study on the relationship between public announcements and stock volatility from Australian Stock Exchange suggests that the non-trading period overnight acts as a barrier for information to be reflected on the stock price. Therefore, the difference between yesterday's closing price and today's opening prices is normally negligible [25]. It can also be seen from Figure 2 that the opening price follows the trace of the closing price really closely. This justifies the above strategy to operate at the opening price instead of the closing price, when the news releases take place during non-trading hours.

If information from both the news releases and past prices are combined together, then the decision follows the predictions made from technical indicators at a day's close and it is revised at the next day's open based on the news predictions (if present). If the two predictions are contradictory, then the outcome of the news always supersedes, as discussed earlier. It operates as follows:

- If news releases are absent, the strategy is unchanged from strategy one above.
- If the share has been bought and the news prediction is negative for that day, the share will be sold at the opening price

- If the share has been bought and the news prediction is positive, do nothing
- If the share has been sold and the news prediction is negative, do nothing.
- If the share has been sold and the news prediction is positive, buy at the opening price.

Table 4 below displays the net compound profit of the system in two month with different information input. A trading system that employs random strategy that has approximately the same number of transactions is used as the benchmark for comparison.

It can be seen that by supplementing conventional technical indicators with the influences of news releases, the proposed system demonstrates promising results under real life situation. Since the stock market is an extremely complicated system, richer information source will be able to provide a better model.

**Table 4.** Compound net profit for different inputs

| System | No. of trades | Net profit |
|---|---|---|
| Random | 8 | $-54 |
| Price Only | 9 | $284 |
| News Only | 7 | $275 |
| Price and News | 11 | $511 |

## 5 Conclusions

In this paper, news paper releases are combined with the technical indicators to predict daily direction of a stock price using SVM. The case study results showed that both the prediction performance and the profitability of the system are enhanced.

However, the current system only categorizes the output into simple rise/fall without specifying the level of change. Therefore, future research efforts will focus on refining the prediction of price trends. Furthermore, the general applicability of the system also needs to be examined further by applying it to other stocks in different sectors.

## References

1. Choi, J.H., Lee, M.K., Rhee, M.W.: Trading S& P 500 Stock Index Futures Using a Neural Network. Proc. of the Annual Int. Conference on Artificial Intelligence Applications on Wall Street, New York (1995) 63–72
2. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
3. Dumais, S., Platt, J., Heckerman, D.: Inductive Learning Algorithms and Representations for Text Categorization. Proc. of the 7th Int. Conf. on Information and Knowledge Management, ACM Press (1998) 148-155

4. Fellbaum, C.(Ed.), WordNet.: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press (1998)
5. Fung, G.P.C., Yu, J.X., Lam, W.: News Sensitive Stock Trend Prediction. Proc. of 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taiwan (2002) 289–296
6. Hellstrom, T., Holmstrom, K.: Predicting the Stock Market. Technical Report Series IMA-TOM-1997-07 (1998)
7. Hotho, A., Stumme, G.: Conceptual Clustering of Text Clusters. Proc. Fachgruppentreffen Maschinelles Lernen Hannover (2002) 37-45
8. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proc. of the 10th European Conference on Machine Learning, Springer, Heidelberg (1998) 137-142
9. Hassan, M.R., Nath, B.: Stock Market Forecasting Using Hidden Markov Model: a new approach. Proc. of 5th Int. Conf.on intelligent systems design and applications (2005)
10. Kim, K., Han, I.: Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index. Expert Syst. Appl. **19** (2) (2000) 125–132
11. Kim, K.J.: Financial Time Series Forecasting Using Support Vector Machines. Neurocomputing **55** (2003) 307-319
12. Rosso, P., Ferretti, E., Jimenez, D., Vidal, V.: Text Categorization and Information Retrieval Using WordNet Senses. CICLing 2004, LNCS **2945** (2004)
13. Mitchell, M.L., Mulherin, J.H.: The Impact of Public Information on the Stock Market. Journal of Finance **49** (3) (1994) 923–950
14. Mittermayer, M.: Forecasting Intraday Stock Price Trends with Text Mining techniques. Proc. of the 37th Hawaii International Conference on System Sciences, Hawaii (2004)
15. Mukherjee, S., Osuna, E., Girosi, F.: Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. Proc. of the IEEE Workshop on Neural Networks for Signal Processing, Amelia Island, FL (1997) 511–520
16. Nofsinger, J.R.: The Impact of Public Information on Investors. Journal of Banking & Finance **25** (2001)1339-1366
17. Quah, T.S., Srinivasan, B.: Improving Returns on Stock Investment through Neural Network Selection. Expert Syst. Appl. **17** (1999) 295–301
18. Romahi, Y., Shen, Q.: Dynamic Financial Forecasting with Automatically Induced Fuzzy Associations. Proc. of the 9th Inter. Conf. on Fuzzy systems (2000) 493-498
19. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
20. Tay, F.E.H., Cao, L.: Application of Support Vector Machines in Financial Time Series Forecasting. Omega **29** (2001) 309–317
21. Thissen, U., Brakel, R. van, Weijer, A.P.de, Melssen, W.J., Buydens, L.M.C.: Using Support Vector Machines for Time Series Prediction. Chemom. Intell. Lab. Syst. **69** (2003) 35–49
22. Vapnik, V.N.: Statistical Learning Theory, Wiley, New York (1998)
23. Brown, G.W., Cliff, M.T.: Investor Sentiment and the Near-Term Stock Market. Journal of Empirical Finance **11** (2004) 1-27
24. Achelis, S.B.: Technical Analysis from A to Z. Probus Publishing, Chicago (1995)
25. Kalev, P.S., Liu, W.M., Pham, P.K., Jarnecic, E.: Public Information Arrival and Volatility of Intraday Stock Returns. Journal of Banking and Finance **28** (2004) 1441-1467

26. Leigh, W., Purvis, R., Ragusa, J.M.: Forecasting the NYSE Composite Index with Technical Analysis, Pattern Recognizer, Neural Network, and Genetic Algorithm: a Case Studying Decision Support. Decision Support Systems **32** (2002) 361–377
27. Lee, R.S.T.: iJADE Stock Advisor: an Intelligent Agent Based Stock Prediction System Using Hybrid RBF Recurrent Network. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans **34** (3) (2004) 421-428
28. Abdullah, M.H.L.b, Ganapathy, V.: Neural Network Ensemble for Financial Trend Prediction. Proc. TENCON **3** (2000) 157-161
29. Abraham, A., Nath, B., Mahanti, P.K.: Hybrid Intelligent Systems for Stock Market Analysis. Proc. of the Inter. Conf. on Computational Science (2001) 337-345