
**Human and Ecological Risk Assessment: An International Journal**
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400879

# Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis

Nan Li[a]; Xun Liang[a]; Xinli Li[a]; Chao Wang[a]; Desheng Dash Wu[bc]
[a] Institute of Computer Science and Technology, Peking University, Beijing, China [b] School of Science and Engineering, Reykjavik University, Reykjavik, Iceland [c] RiskLab, University of Toronto, Toronto, Canada

Online publication date: 12 February 2010

**To cite this Article** Li, Nan , Liang, Xun , Li, Xinli , Wang, Chao and Wu, Desheng Dash(2009) 'Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis', Human and Ecological Risk Assessment: An International Journal, 15: 2, 227 — 252
**To link to this Article: DOI:** 10.1080/10807030902761056
**URL:** http://dx.doi.org/10.1080/10807030902761056

Human and Ecological Risk Assessment, 15: 227–252, 2009

ISSN: 1080-7039 print / 1549-7680 online
DOI: 10.1080/10807030902761056

# Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis

**Nan Li,[1] Xun Liang,[1] Xinli Li,[1] Chao Wang,[1] and Desheng Dash Wu[2,3]**
[1]Institute of Computer Science and Technology, Peking University, Beijing, China;
[2]School of Science and Engineering, Reykjavik University, Reykjavik, Iceland;
[3]RiskLab, University of Toronto, Toronto, Canada



## ABSTRACT

Under the network environment, the trading volume and asset price of a financial commodity or instrument are affected by various complicated factors. Machine learning and sentiment analysis provide powerful tools to collect a great deal of data from the website and retrieve useful information for effectively forecasting financial risk of associated companies. This article studies trading volume and asset price risk when sentimental financial information data are available using both sentiment analysis and popular machine learning approaches: artificial neural network (ANN) and support vector machine (SVM). Nonlinear GARCH-based mining models are developed by integrating GARCH (generalized autoregressive conditional heteroskedasticity) theory and ANN and SVM. Empirical studies in the U.S. stock market show that the proposed approach achieves favorable forecast performances. GARCH-based SVM outperforms GARCH-based ANN for volatility forecast, whereas GARCH-based ANN achieves a better forecast result for the volatility trend. Results also indicate a strong correlation between information sentiment and both trading volume and asset price volatility.

**Key Words:** risk, natural language processing, sentiment analysis, stock market, machine learning.

## INTRODUCTION

In today's global economy, effective forecast of financial risk through typical financial measures has been considerably appealing. Meanwhile, it is also a very challenging task to various financial practitioners due to a more and more complicated environment business institutions are facing. A typical term reflecting the movement is the financial volatility, which is a required parameter for pricing many kinds of financial assets and derivatives. Additionally, it is well acknowledged that financial volatility implies financial risk. Therefore, accurate prediction of financial volatility is of critical significance. Nonetheless, how to efficiently predict financial

Address correspondence to Desheng Dash Wu, School of Science and Engineering, Reykjavik University, Kringlunni 1, IS-103 Reykjavik, Iceland. E-mail: dash@ru.is

volatility has been one of the unsolved issues under investigation. In this article, we aim to establish a scalable and customizable mathematical model to achieve this goal.

A great deal of information from the network can influence stock market movement. Widely researched factors include consumable prices, interest rates, foreign exchange rates, and so on. Financial information from the network has not attracted enough attention, although some researchers observed that financial information has become increasingly influential (Chuttur and Bhurtun 2005; Costantino *et al.* 1997). A lot of financial information is available from the Internet in this information era. In this regard, two questions are naturally raised: first, is there an obvious association or correlation between the online financial news and the financial risk measured in volatility? Second, given that an obvious association or correlation exists, how can we use this correlation relationship to predict financial volatility? Some researchers seek to mine the correlations between stock markets and the frequencies of financial keywords occurring in news, or focus on public resolutions, reports, and information disclosure of listed companies (Ettredge *et al.* 2000; Freisleben and Ripper 1997). No research has considered both the information volume and the information sentiment, as was done in our study.

In this article, we investigate the correlations between the financial volatility, that is, asset price volatility and trading volume volatility, and the financial information, by developing two GARCH-based (Bollerslev 1986) data mining models: GARCH-based artificial neural network (ANN) (Cai and Shi 2003; Huang *et al.* 2006) and GARCH-based support vector machine (SVM) (Cristianini and Shawe-Taylor 2000; Vapnik 1998) approaches.

The GARCH (generalized autoregressive conditional heteroskedasticity) model has been widely applied to model financial time series investigation due to its properties of being able to capture various financial features such as fat tail and asymmetry and of simple implementation (Sohn and Lim 2007). ANN and SVM are two popular data mining approaches widely applied for financial computation (Wu *et al.* 2006; Freisleben and Ripper 1997; Tino *et al.* 2001; Trafalis and Ince 2000; Zhang *et al.* 2003). Using GARCH-based ANN and SVM allows us to dynamically identify and record association information mined from on-line financial information, which will then be used to forecast financial volatility. Sentiment analysis is employed to probe into the correlations between information sentiment and asset price volatility. This is a text mining technique employed by a number of researchers to determine the attitude of a speaker or a writer with respect to a specific topic (Ahmad and Almas 2005; Chaovalit and Zhou 2005; Turney 2001, 2002; Turney and Littman 2003). Results are compared using innovative statistic measures.

The rest of this article is organized as follows. The next section briefly explains the architecture of our overall approaches, following which a section details the time series models and volatility calculation. How to utilize ANN and SVM to implement a dynamic training and forecasting for mining correlations between information and volatility is presented in the next two sections, followed by empirical studies.

## MODELS

As aforementioned, we investigated the associations between financial volatility and information via two steps. During the first step, machine learning models

utilizing both ANN and SVM were established to probabilistically model the correlation between information volume and trading volume volatility. In the second step, we used the sentiment analysis and incorporated the emotional polarity (positive or negative) of the authors to pinpoint financial news analysis. The GARCH model was used to develop GARCH-based ANN and SVM.

### Information Volume and Volatility

It is recognized that the fluctuation of trading volume performs like stock price fluctuation and vividly reflects the market behavior. We thus probe into the associations between the trading volume volatility and the online information volume. Online financial information volume has been assumed as an important element that affects financial trading volume volatility. We forecast the volatility, partly relying on the online information, using both an ANN-based and SVM-based approach. Eventually, we conduct a comparison between these two to observe the different forecast performances. The basic architecture of this approach can be visualized in Figure 1.

We downloaded the online financial information from Google Finance (http://www.finance.google.com) as shown in Figure 1, and thereafter post-processed the data to acquire the daily information volumes for various stocks and indices. In Table 1 is shown a snippet of the post-processing result, with each sub-row on top indicating the date and the one below the volume of the news. In Table 1, N, D, M, I, A, and G stand for NASDAQ, DOW, MSFT, INTC, AAPL and GOOG, respectively; 0629 stands for June 29, 2006.

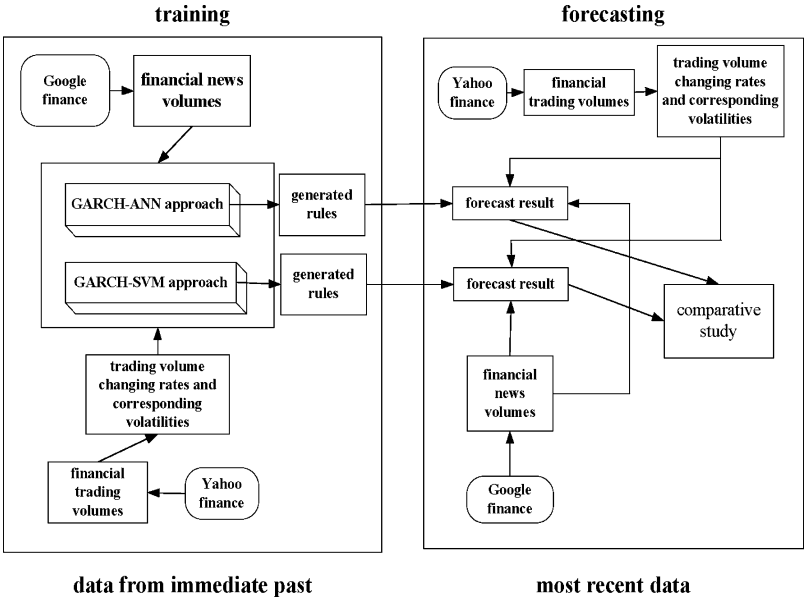


**Figure 1.** Flow chart and functional parts of our approach to associate information volume and volatility.

**Table 1.** The information volumes calculated from Google Finance.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| N | 0629 | 0630 | 0703 | 0704 | 0705 | 0706 | 0707 | ... |
| | 7 | 4 | 5 | 2 | 4 | 4 | 3 | ... |
| D | 0630 | 0703 | 0704 | 0705 | 0706 | 0707 | 0708 | ... |
| | 2 | 3 | 3 | 4 | 5 | 12 | 1 | ... |
| M | 0628 | 0629 | 0630 | 0701 | 0703 | 0704 | 0705 | ... |
| | 5 | 8 | 3 | 1 | 2 | 2 | 4 | ... |
| I | 0703 | 0705 | 0706 | 0707 | 0710 | 0711 | 0712 | ... |
| | 2 | 2 | 2 | 3 | 2 | 1 | 2 | ... |
| A | 0630 | 0701 | 0703 | 0705 | 0706 | 0707 | 0710 | ... |
| | 7 | 1 | 4 | 3 | 6 | 2 | 2 | ... |
| G | 0717 | 0718 | 0719 | 0720 | 0721 | 0724 | 0725 | ... |
| | 3 | 5 | 2 | 8 | 7 | 4 | 5 | ... |



**Information Sentiment and Volatility**

Representing financial information purely by its volume might be misleading, and thus undermine the efficiency of the forecast. In order to investigate the impacts of on-line information upon financial time series in a comprehensive manner, we compensate the deficiency by exploring its content as well; specifically, its emotional or sentimental polarity. One essential step is the sentiment analysis for each news entry, which is primarily accomplished by a bags-of-words–based methodology. We then obtain a real value for each news entry, with the sign being the authors' judgmental state and the absolute value of how intense the emotion is.

HowNet (http://www.keenage.com/html/e_index.html), an on-line commonsense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoted in lexicons of the Chinese and their English equivalents, is utilized to approach the sentiment calculation for the whole piece of news via the keywords. Each piece of news is decomposed and converted into a keyword array, with the same sequence as in the article, each of which is assigned a specific sentiment value based on the HowNet word. The overall sentiment for the whole article is acquired by combining all the sentiment values of those keywords. After the sentiment time series is obtained, it will be fed into the machine learning system, SVM in particular here, as one of the exogenous inputs. By assigning sentiment as one element of the feature vector for a listed company, nonlinear correlation between news sentiment on-line and financial volatility will be quantitatively explored. The basic architecture of this approach can be visualized in Figure 2.

Financial news used in this phase is acquired from a variety of online sources, and experiments are carried on a huge body of listed companies in the U.S. stock markets. Aggregated statistical results are the key part of the empirical studies to substantiate the nonlinear correlation between the two entities. Shown in Table 2 is a snippet of the financial news entries with their calculated sentiment values, where 2007-1 indicates this is a news entry in the first week of 2007.

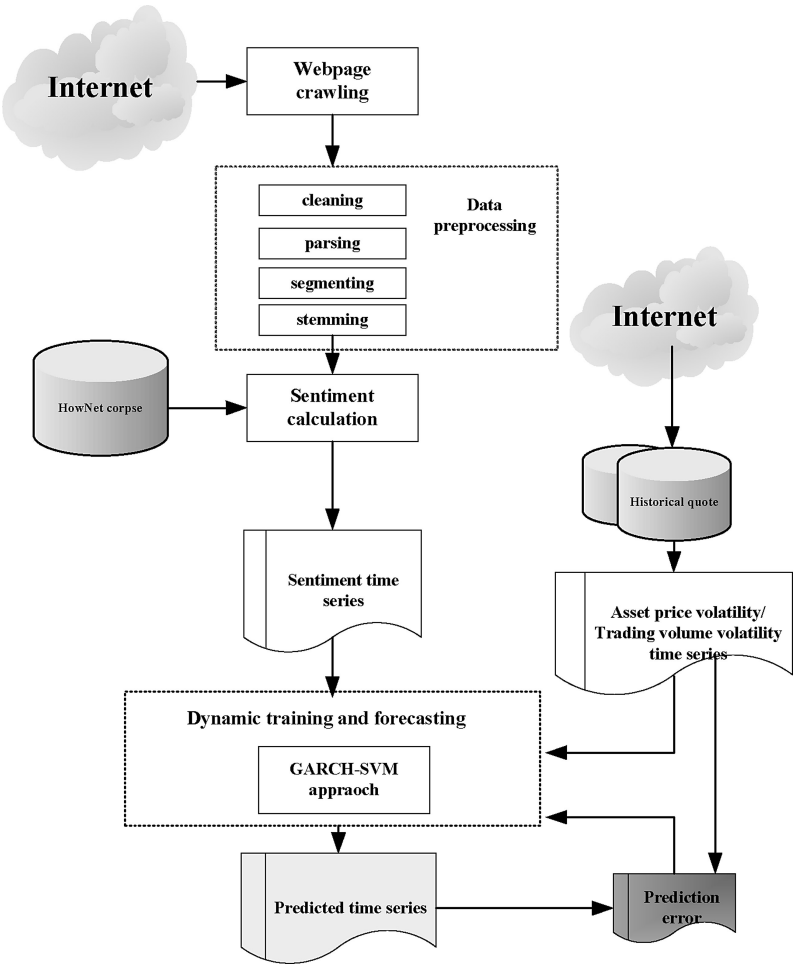


**Figure 2.** Flow chart to associate information sentiment and volatility.

## VOLATILITY MODEL AND MODIFIED NONLINEAR GARCH MODEL

Volatility refers to the standard deviation or variance of the change in value of a financial instrument within a specific time span. The GARCH system is widely employed in modeling financial time series that exhibit time-varying volatility clustering. In this section, we develop a GARCH system by incorporating financial information into the usual framework.

### Modified Nonlinear GARCH Model

The GARCH model, proposed by Bollerslev in 1986, can be formulated as

$$y_t = \mu_t + \varepsilon_t, \tag{1}$$

$$\varepsilon_t | \psi_{t-1} \sim N\left(0, \sigma_t^2\right), \tag{2}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}^2, \ (\alpha_0 > 0, \alpha_i, \ \beta_i \geq 0) \tag{3}$$

**Table 2.** A snippet of news entries for the companies ADCT, S, and MRO.

| News ID | News title | Time window | Company symbol | News body | News sentiment |
|---|---|---|---|---|---|
| 2007010102057465 | An Insecure Future for McAfee | 2007-1 | ADCT | Perhaps it's fatigue with the options scandal that has now spread to more than 100 technology companies. Perhaps it's . . . | 45.4 |
| 2007010202063354 | Stocks end 2006 with best gains in 3 years | 2007-1 | S | NEW YORK (MarketWatch)— U.S. stocks finished the year with strong gains Friday, with all three major stock averages booking their best performance since 2003. The Dow Jones Industrial Average ($INDU :$INDUNews | 17.8 |
| 2007050203261590 | Marathon Oil's lower earnings top forecasts | 2007-18 | MRO | SAN FRANCISCO (MarketWatch)— Marathon Oil Corp. reported Tuesday a drop in first-quarter earnings, clipped by lower oil and gas prices and a decline in production. For the three months ended March 31, Marathon (MRO: MRONews | −0.8 |
| 2007060203595695 | Still looking good | 2007-22 | ADCT | ANNANDALE, Va. (MarketWatch)— It's been a little bit over two months since the triggering of a rare, and historically very bullish, technical signal. (Read March 22 column.) Can we count on the bullish winds of that signal blowing into the. . . | 11.8 |

where the daily return $y_t$ is sum of the deterministic mean return $\mu_t$ and a stochastic term $\varepsilon_t$, also known as the shock, forecast error, residual, innovation, and so on (Burges 1998; Freisleben and Ripper 1997), $\psi_{t-1}$ represents the information set available at time $t$ and $\sigma_t^2$ is the time-varying variance of both $y_t$ and $\varepsilon_t$. In our approach, we substitute $y_t$ with the daily changing rate of either the trading volume or the asset price.

The GARCH model has indicated that $\varepsilon_t$ is a function of those exogenous inputs, which somewhat affects the financial volatility. The GARCH model bases its conditional distribution on the information set available at time $t$. Freisleben and Ripper (1997) point out that the parameter $\beta_i$ in Eq. (3) describes the stock return's immediate reaction to new events in the market, mostly in the form of financial news (Freisleben and Ripper 1997). In the meanwhile, the fast development of the Internet enables us to acquire the online financial information in a most real-time and exhaustive fashion. Considering these factors, to designate the financial information volume as one variate of $\varepsilon_t$ is justifiable.

Therefore we formulize $\varepsilon_t$ using the following two equations,

$$\varepsilon_t = y_t - \zeta, \tag{4}$$

$$\varepsilon_t = f_t(W_t, \varepsilon_t') = g_t(W_t) + \theta_t \varepsilon_t', \tag{5}$$

where $\zeta$ is a constant and $W_t$ is the online financial information volume on day $t$. Consequently a modified GARCH model can be expressed as

$$y_t = \mu_t + \varepsilon_t, \tag{6}$$

$$\varepsilon_t | \psi_{t-1} \sim N\left(0, \sigma_t^2\right), \tag{7}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \chi_{t-i}\left(\sigma_{t-i}^2\right) + \sum_{j=1}^{q} \beta_j \varphi_{t-j}\left(y_{t-j}^2\right) + \sum_{k=1}^{r} \gamma_k \phi_{t-k}\left(W_{t-k}^2\right) \tag{8}$$

where $p$, $q$, $r$represent the three time lags, the three unknown functions, $\chi_{t-i}$, $\varphi_{t-j}$, and $\phi_{t-k}$ represent the undetermined nonlinear correlations.

Financial time series exhibit specific features (Freisleben and Ripper 1997), which renders the GARCH model a preferable alternative over other counterparts. In Figure 3 are depicted the daily changing rates of trading volumes of the NASDAQ index within the period from October 11, 1984 to October 16, 2006, which exhibits obvious volatility-clustering feature. Besides, a Kurtosis value of 11.53 calculated from it also implies that there exists an underlying fat tail effect. In parallel, we have discovered obvious GARCH effects exhibited by on-line financial information volume time series as well, via tests conducted on more than 100 stocks in the U.S. stock markets. Therefore, it is relevant to assume that the volatility of financial trading volume exhibits similar characteristics of stock price, that is, GARCH effects.

### Daily Volatility Model

For a specific stock or index, let $v_t$ denote its trading volume on day $t$, the daily changing rate $y_t$ of the trading volume is denoted as

$$y_t = \ln \frac{v_t}{v_{t-1}}. \tag{9}$$

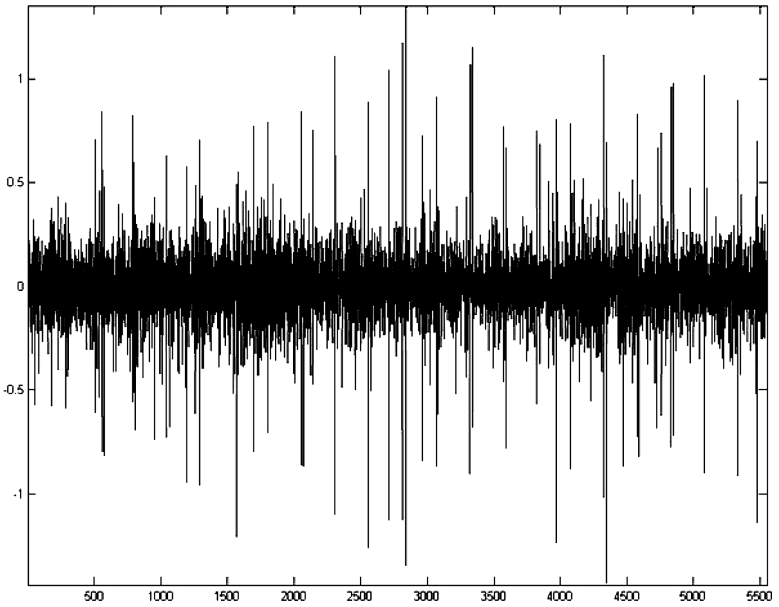


**Figure 3.** Daily changing rates of the trading volumes of the NASDAQ index within the period from October 11, 1984 to October 16, 2006.

If $D$ is defined as the width of the calculating window, the volatility $\sigma_t^2$ can be obtained by computing the variance of the $y_t$ within the $(t - D + 1)$th and the $t$th day as

$$\sigma_t^2 = \frac{\sum_{i=0}^{D-1} (y_{t-i} - \overline{y_t})^2}{D-1}, \tag{10}$$

where

$$\overline{y_t} = \frac{\sum_{i=0}^{D-1} y_{t-i}}{D}. \tag{11}$$

Such daily volatilities are calculated based on a sliding volatility window of a certain length. Each day's volatility represents the variation in the value of trading volume over the past several days until the current day.

### Volatility Model in Time Windows

This section discusses time-window based volatility models. A time-window based volatility model enables the decision-maker to get a clearer view of the volatility movement in a more macro-scopical fashion and especially useful in mining correlations between information sentiment and volatility.

This method starts with a time segmentation step. Suppose there is no trading activity during weekends in stock markets, a time window of one week is used to decompose the whole time span into several consecutive time windows, each with the same length of one week shown in Figure 4. In Figure 4, $N$ is the total number

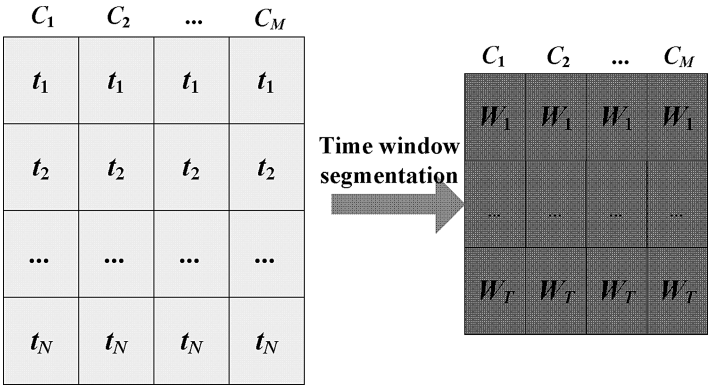


**Figure 4.** Date unification for various listed companies.



of trading days for all companies, while $T$ is the total number of time windows. Suppose that the length of a window is $L$, then we obtain $N = T^*L$. Due to the fact that various listed companies have a diverse span of trading dates, date unification constitutes a necessary step, which is visualized in Figure 5, where there are $M$ listed companies in total, with $N_j$ (j = 1, 2, . . . , $M$) representing the total number of trading days within a predetermined time horizon for each. The segmentation and unification processes are applied upon the quotes of all these listed companies.

Let $W_i$ denote the current time window. For each trading day $t$ within $W_i$, the closing price and the trading volume are denoted as $p_t$ and $v_t$, respectively. There are $L$ days in this window and the starting day is denoted as $t_i$. Therefore the volatility

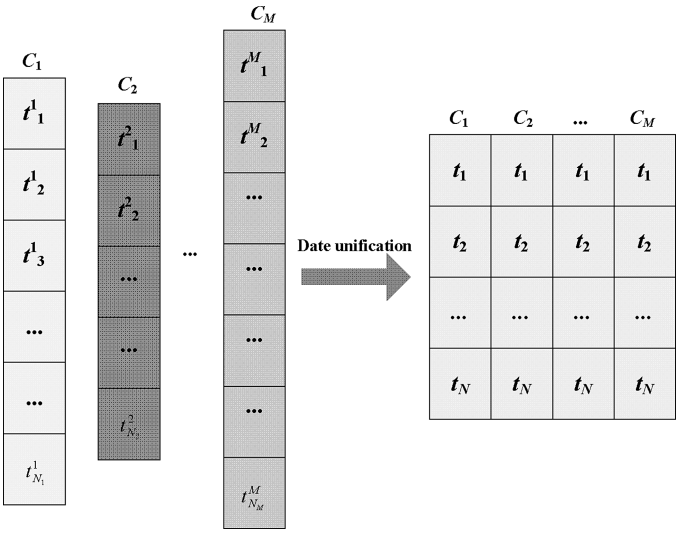


**Figure 5.** A time window segmentation process.

of asset price within this window can be formulated as

$$\sigma_i^{p2} = \frac{\sum_{t=t_i}^{t_i+L-1} (y_t^p - \overline{y_t^p})^2}{L-1}, \tag{12}$$

where $y_t^p$ represents the changing rate of asset price on day $t$, $L$ is the length of the time window and

$$\overline{y_i^p} = \frac{\sum_{t=t_i}^{t_i+L-1} y_t^p}{L}. \tag{13}$$

Similar formulations could be applied upon trading volume as well, thus we have

$$\sigma_i^{v2} = \frac{\sum_{t=t_i}^{t_i+L-1} (y_t^v - \overline{y_t^v})^2}{L-1}, \tag{14}$$

as the trading volume volatility within $W_i$ where

$$\overline{y_i^v} = \frac{\sum_{t=t_i}^{t_i+L-1} y_t^v}{L}. \tag{15}$$

## USING ANN AND SVM TO ASSOCIATE INFORMATION VOLUME AND TRADING VOLUME VOLATILITY

Machine learning is the key technique to establish a bridge between information volume and trading volume volatility time series. By training the learning models based on training samples formatted from actual data sets in a progress fashion on sliding training windows, new predictions for the future could be performed. The volatility in this study will be calculated according to the daily volatility model introduced in the Daily Volatility Model Section.

### GARCH-Based ANN

As shown in Figure 6, the proposed the GARCH-based ANN is a three-layer feed-forward neural network where $\hat{\sigma}_t^2$ represents the forecast volatility on day $t$ , $y_t^2$ denotes the squared daily changing rate of the trading volume, and $W_t^2$ represents the squared daily financial information volume. The input elements $\sigma_{t-1}^2$ to $\sigma_{t-p}^2$ constitute the auto-regressive part of the model whereas the elements $y_{t-1}^2$ to $W_{t-r}^2$ represent the moving average part. The size of the input vector amounts to $p+q+r$; there are $H$ units in the hidden layer and the size of the output vector is 1. The training algorithms used in our approach consist of the Bayesian regularization back-propagation, BFGS quasi-Newton back-propagation (Hu *et al.* 2006), and the Levenberg-Marquardt back-propagation (Kanzow *et al.* 2004), with the former two able to effectively impede the over-fitting.

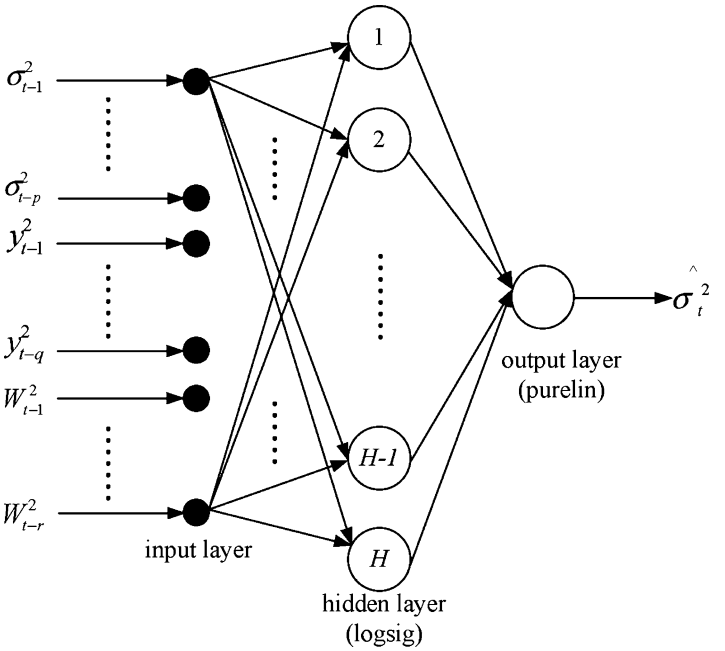


**Figure 6.** The GARCH-based ANN architecture.



### GARCH-Based SVM

The SVM is based on statistical learning theory, and the training of SVM leads to a quadratic programming (QP) problem. Proposed by Vapnik (1998), SVM has become a promising data mining technique to overcome problems such as over-fitting and local minimum while providing high generalization (Zhang *et al.* 2003). SVM has been widely applied in various fields such as classification (Burges 1998; Zhang *et al.* 2003; Eitrich and Lang 2006), regression (Wang and Hu 2005), image retrieval (Tong and Chang 2001), text categorization, time series analysis (Wu *et al.* 2004), and so on. It is worth mentioning that SVM has also been adopted by many researchers in the financial field to forecast financial time series (Trafalis and Ince 2000).

In Figure 7 is shown the GARCH-based SVM model we employed to realize function approximation and regression estimation, where all variables and parameters are defined in Figure 6. The basic idea of regression SVM is to nonlinearly map the input training data via mapping function into a higher dimensional feature space and then a linear regression problem is obtained and solved in this feature space (Cristianini and Shawe-Taylor 2000; Vapnik 1998).

### Dynamic Training Technique

Because financial data are inherently characteristic of being noisy, stochastic, and non-stationary, we adopt a special training process called "dynamic training technique" to prevent over-fitting in both the GARCH-based ANN and the GARCH-based SVM. To implement this technique within one certain training cycle, if we

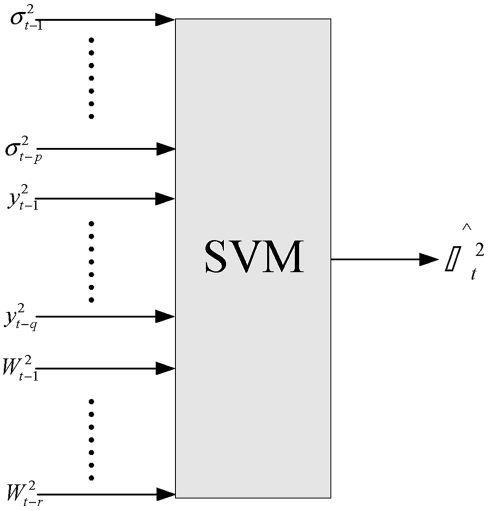


**Figure 7.** The GARCH-based SVM architecture.

want to forecast the volatility on day $t$, provided that the size of the training samples amounts to $C$, we first train an ANN or SVM via supervised predictions for the volatilities from day $t-C$ to day $t-1$. Thereafter, forecast for the volatility on day $t$ is carried out using the trained ANN or SVM. In a nutshell, the dynamic training refers to an iterative process within which the forecasting for the volatility on a certain day always utilizes a newly trained ANN or SVM trained by data attained in the most immediate past.

To further demonstrate the dynamic training, let $C$ denote the size of the training samples for one training cycle. In addition, we use a matrix tuple $\langle P, T\rangle$ to denote the training samples for one training cycle; thus we have

$$P = \begin{bmatrix} \sigma^2_{t-C-1} & \sigma^2_{t-C} & \sigma^2_{t-C+1} & \cdots & \sigma^2_{t-2} \\ \sigma^2_{t-C-2} & \sigma^2_{t-C-1} & \sigma^2_{t-C} & \cdots & \sigma^2_{t-3} \\ \cdots & & & & \\ \sigma^2_{t-C-p} & \sigma^2_{t-C-p+1} & \sigma^2_{t-C-p+2} & \cdots & \sigma^2_{t-1-p} \\ y^2_{t-C-1} & y^2_{t-C} & y^2_{t-C+1} & \cdots & y^2_{t-2} \\ y^2_{t-C-2} & y^2_{t-C-1} & y^2_{t-C} & \cdots & y^2_{t-3} \\ \cdots & & & & \\ y^2_{t-C-q} & y^2_{t-C-q+1} & y^2_{t-C-q+2} & \cdots & y^2_{t-1-q} \\ W^2_{t-C-1} & W^2_{t-C} & W^2_{t-C+1} & \cdots & W^2_{t-2} \\ W^2_{t-C-2} & W^2_{t-C-1} & W^2_{t-C} & \cdots & W^2_{t-3} \\ \cdots & & & & \\ W^2_{t-C-r} & W^2_{t-C-r+1} & W^2_{t-C-r+2} & \cdots & W^2_{t-1-r} \end{bmatrix} \tag{16}$$

and

$$T = \left[\sigma_{t-C}^2 \quad \sigma_{t-C+1}^2 \quad \sigma_{t-C+2}^2 \quad \cdots \quad \sigma_{t-1}^2\right] \tag{17}$$

The trained ANN or SVM is then used to forecast the volatility on the day $t$, which is referred to as the testing process. Likewise, we denote the testing samples by another matrix tuple $\langle P', T' \rangle$ as

$$P' = \begin{bmatrix} \sigma_{t-1}^2 \\ \sigma_{t-2}^2 \\ \cdots \\ \sigma_{t-p}^2 \\ y_{t-1}^2 \\ y_{t-2}^2 \\ \cdots \\ y_{t-q}^2 \\ W_{t-1}^2 \\ W_{t-2}^2 \\ \cdots \\ W_{t-r}^2 \end{bmatrix}, \tag{18}$$

and

$$T' = \left[\sigma_t^2\right], \tag{19}$$

where $T'$ stands for the real volatility on the day $t$.

## USING GARCH-BASED SVM TO ASSOCIATE INFORMATION SENTIMENT AND ASSET PRICE VOLATILITY

This section investigates the relationship between volatility and information sentiment. We incorporate the average sentiment value into the current forecasting model in the previous section and predict how the volatility will move in the immediate future from a more comprehensive perspective. The volatility in this phase of study will be calculated by the volatility model in time windows introduced previously.

### Sentiment Analysis for Financial News

Sentiment analysis for an article usually includes two types of approaches, semantic orientation based approach (Chaovalit and Zhou 2005; Turney 2001, 2002; Turney and Littman 2003) and machine learning–based approach (Chaovalit and Zhou 2005). Researchers have conducted comparative studies to those two methodologies for movie reviews (Chaovalit and Zhou 2005). Considering the number of articles that will be analyzed in our study, which renders having people label the training

**Table 3.** The eight word sets we use in this article to calculate the keyword sentiment.

| Word sets | Description |
|---|---|
| POSITIVE | A list of English words that have positive emotional polarity, which includes a set of 4363 words |
| NEGATIVE | A list of English words that have negative emotional polarity, which includes a set of 4574 words |
| PRIVATIVE | A list of privative English words, which includes 14 words. {no, not, none, neither, never, hardly, seldom, barely, scarcely, ain't, aren't, isn't, hasn't, haven't} |
| | The following five sets are modifiers, whose intensities decrease while $i$ increases. |
| $\text{MODIFIER}_1$ | 64 modifier words, with $\text{WEIGHT}_1 = 2$ |
| $\text{MODIFIER}_2$ | 25 modifier words, with $\text{WEIGHT}_2 = 1.8$ |
| $\text{MODIFIER}_3$ | 22 modifier words, with $\text{WEIGHT}_3 = 1.6$ |
| $\text{MODIFIER}_4$ | 15 modifier words, with $\text{WEIGHT}_4 = 1.4$ |
| $\text{MODIFIER}_5$ | 11 modifier words, with $\text{WEIGHT}_5 = 0.8$ |



samples way too time-consuming, we decided to take the semantic orientation–based methodology.

In this article, the sentiment of the whole piece of news is based on the sentiment values of all its keywords, that is, those contained in the article with emotional polarity. The English lexicon released by HowNet is used in our article to form the essential word sets based on which sentiment calculation for a keyword is implemented. Eventually, we generate eight word sets: POSITIVE, NEGATIVE, PRIVATIVE, $\text{MODIFIER}_i (i = 1, 2, 3, 4, 5)$. Each $\text{MODIFIER}_i$ is bound with a weight value $\text{WEIGHT}_i$, which denotes the intensity of the words listed in this set. The intensity of the words increases with the weight value. The definition for these eight sets is given in Table 3.

Keyword sentiment is calculated based on looking into those word sets and counting the matches, according to a designed algorithm with flowchart (shown in Figure 8). The sentiment for the whole article is calculated by adding up the sentiments for all the keywords that are contained.

### Using GARCH-Based SVM to Associate Information Sentiment and Volatility

As mentioned at the beginning of this section, the volatility will be calculated on a time window basis, which means the time series will be first segmented into a series of time windows. The GARCH-based SVM approach is performed on a sliding window basis. Each SVM will be trained by data collected from the current and the previous time window. The well-trained SVM is used to predict volatility in the next time window. In Figure 9 is visualized the process of sliding time window machine learning, where $W_{i-1}$ and $W_i$ constitute the training input and output, respectively, and $W_i$ and $W_{i+1}$ constitute the forecasting input and output, respectively.

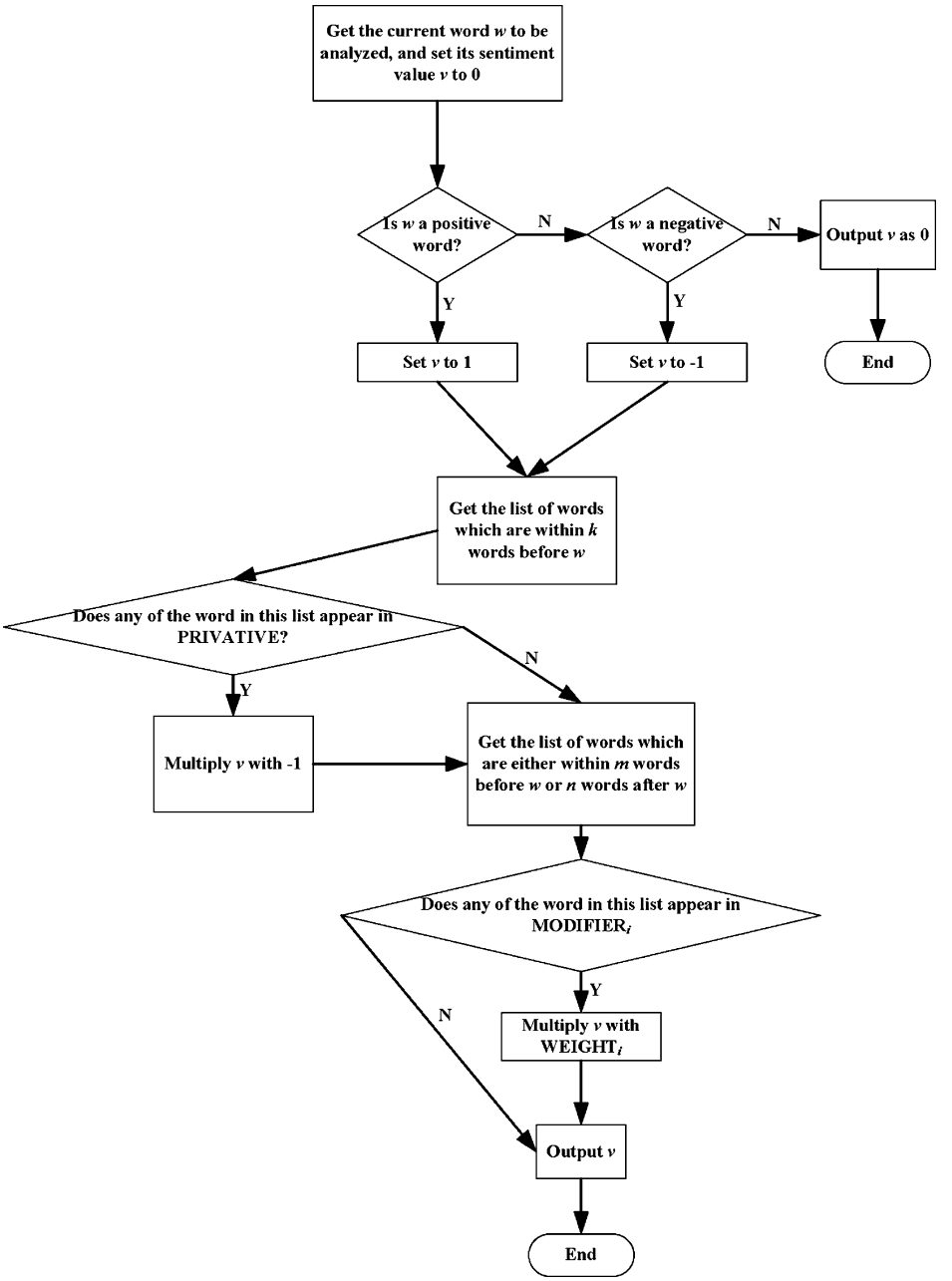


**Figure 8.** Sentiment calculation process for the current keyword $w$.

Each time window in Figure 9 corresponds to an input and output matrix formatted for machine learning and predicting. These matrices contain expanded companies in order to generate aggregated statistic values of the whole stock market. Suppose there are $M$ listed companies of interests, let $< I, O >$ denote the training input and output matrix tuple to forecast the volatility in time window $W_i$,

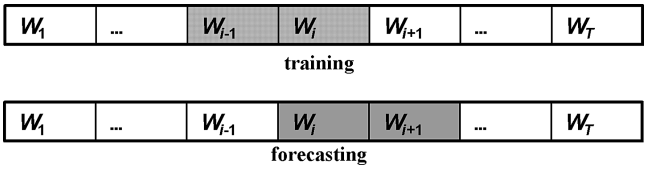


**Figure 9.** Sling time window learning and forecasting.

thus we have

$$I = \begin{bmatrix} \sigma_{i-1}^{p2}(1) & \bar{y}_{i-1}^{p^2}(1) & S_{i-1}^2(1) \\ \sigma_{i-1}^{p2}(2) & \bar{y}_{i-1}^{p^2}(2) & S_{i-1}^2(2) \\ \cdots & & \\ \sigma_{i-1}^{p2}(M) & \bar{y}_{i-1}^{p^2}(M) & S_{i-1}^2(M) \end{bmatrix} \tag{20}$$

and

$$O = \begin{bmatrix} \sigma_i^{p2}(1) \\ \sigma_i^{p2}(2) \\ \cdots \\ \sigma_i^{p2}(M) \end{bmatrix}, \tag{21}$$

where $\sigma_{i-1}^{p2}(k)$ $(k = 1,2,3,\ldots M)$ is the asset price volatility within the time window $W_{i-1}$ for the company $k$, $\bar{y}_{i-1}^{p^2}(k)$ represents the average daily changing rate of the asset price of company $k$ in $W_{i-1}$, and $S_{i-1}^2(k)$ is the sum of the sentiment values for all the news entries relating to company $k$ within the time window $W_{i-1}$.

Accordingly, denote by $\langle I', O' \rangle$ the forecasting input and output matrix tuple for $W_i$, we have

$$I' = \begin{bmatrix} \sigma_i^{p2}(1) & \bar{y}_i^{p^2}(1) & S_i^2(1) \\ \sigma_i^{p2}(2) & \bar{y}_i^{p^2}(2) & S_i^2(2) \\ \cdots & & \\ \sigma_i^{p2}(M) & \bar{y}_i^{p^2}(M) & S_i^2(M) \end{bmatrix} \tag{22}$$

and

$$O' = \begin{bmatrix} \sigma_{i+1}^{p2}(1) \\ \sigma_{i+1}^{p2}(2) \\ \cdots \\ \sigma_{i+1}^{p2}(M) \end{bmatrix}. \tag{23}$$

## EMPIRICAL RESULTS AND ANALYSIS

The empirical studies are roughly composed of two parts. In the first experiment, we utilize both the GARCH-based ANN and SVM to study the correlations between financial information volume and trading volume volatility, and conduct a comparative study of different machine learning techniques in financial volatility forecasting. A daily volatility model serves as the main volatility calculating approach in this experiment. The datasets in this step are limited to the trading quote and news data for two indices and two listed companies in the U.S. stock markets. The time horizon spans across 3 months. In the second experiment, we apply the GARCH-based SVM model to data of the whole 2007 year for 177 listed companies in the U.S. markets. Time window based volatility model is used instead in this experiment. For both experiments, historical financial quotation data is downloaded and formatted from Yahoo Finance (http://www.finance.yahoo.com).



### Trading Volume Volatility Forecasting

#### Experiment design and results

We utilize both ANN and SVM to carry out the forecast for the volatilities. The SVM toolbox adopted in this article is the LS-SVMlab (http://www.esat.kuleuven.ac.be/sista/lssvmlab/). The least squares SVM (LS-SVM) (Wang and Hu 2005) is a reformulation of the regular SVM. Besides, the kernel function selected in our approach is the RBF function. Only 15 Web pages are downloaded from a constant set of news sources on the morning of the current day, whereas in our approach, we based our experiments on the financial information acquired from Google Finance, yielding a comprehensive set of on-line financial information within the past 3 months from more than 500 financial portals on-line. The widths of the volatility calculating windows, namely the $D$, in both phases are set to 20 days. Here we further introduce the concept of volatility trend, which is denoted as $\Delta\sigma_t^2$. This item indicates the moving direction of the daily volatility, which can be formulated as:

$$\Delta\sigma_t^2 = \begin{cases} 1, & \sigma_t^2 > \sigma_{t-1}^2 \\ 0, & \sigma_t^2 = \sigma_{t-1}^2 \\ -1, & \sigma_t^2 < \sigma_{t-1}^2 \end{cases}, \tag{24}$$

where $\sigma_{t-1}^2$ represents the daily trading volume volatility on the day $t-1$.

We use two criteria to measure the forecasting performance: the average forecast error $\bar{e}$ and the volatility trend forecast accuracy *ratio*. Let $e_t$ denote the forecast error on the day $t$, thus we have

$$e_t = \frac{|\hat{\sigma}_t^2 - \sigma_t^2|}{\sigma_t^2}, \tag{25}$$

$$\bar{e} = \frac{1}{t_1 - t_0 + 1} \sum_{t=t_0}^{t_1} e_t, \tag{26}$$

where $t_0$ and $t_1$ represent the beginning and ending day for the forecasting.

Suppose that we use $\hat{\sigma}_t^2$ to denote the forecasted volatility on day $t$, if $\hat{\sigma}_t^2 - \sigma_{t-1}^2$ and $\sigma_t^2 - \Delta\sigma_{t-1}^2$ share the same sign, we say that an accurate forecast for the volatility trend $\Delta\sigma_t^2$ has been achieved. The aforementioned *ratio* is defined as the percentage of the days on which the forecast volatility trend has been accurately forecasted.

We conduct our experiments on two indices, NASDAQ and DOW, and two stocks, MSFT and INTC, with the time span as from June 30, 2006 to September 28, 2006, from June 29, 2006 to September 26, 2006, from June 28, 2006 to September 26, 2006, and from July 3, 2006 to September 28, 2006, respectively. The optimal parameters for RBF in SVM, namely the regularization parameter *gam* and the bandwidth *sig2*, are set as $sig2 = 50$ and $gam = 10$. The results of these experiments are shown in Table 4, where s/i stand for stock/index. N, D, M, I stand for NASDAQ, DOW, MSFT, INTC, respectively. $C$ represents the size of the samples, $H$ the number of the hidden nodes, and $p$, $q$, $r$ the 3 time lags. In order to compare different results, we alter the parameter values for both GARCH-based ANN and SVM three times. In Table 4 are given the predicted values of the average forecast error and the volatility trend forecast accuracy *ratio* for these three scenarios. Results indicate that the GARCH-based SVM outperforms the GARCH-based ANN for *volatility* forecast, whereas the GARCH-based ANN achieves a better forecast result for the *volatility trend*. This is primarily because SVM is characteristic of the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin, which demonstrate its better generality in overcoming over-fitting phenomenon. In addition, SVM is a preferable solution, especially for

**Table 4.** Predicted values of the average forecast error and the volatility trend forecast accuracy *ratio*.

| s/i | Model | $C$ | $H$ | $p$ | $q$ | $r$ | $\bar{e}$(%) | *Ratio* (%) |
|---|---|---|---|---|---|---|---|---|
| N | ANN | 20 | 5 | 4 | 9 | 1 | 11.08 | 83.33 |
| | SVM | 20 | — | 4 | 9 | 1 | 9.60 | 58.33 |
| | ANN | 10 | 5 | 5 | 8 | 2 | 9.33 | 73.91 |
| | SVM | 10 | — | 5 | 8 | 2 | 7.30 | 52.17 |
| | ANN | 10 | 4 | 3 | 8 | 3 | 7.67 | 69.57 |
| | SVM | 10 | — | 3 | 8 | 3 | 7.62 | 47.83 |
| D | ANN | 10 | 8 | 6 | 6 | 1 | 10.17 | 68.00 |
| | SVM | 10 | — | 6 | 6 | 1 | 9.42 | 64.00 |
| | ANN | 15 | 5 | 9 | 4 | 2 | 12.96 | 64.71 |
| | SVM | 15 | — | 9 | 4 | 2 | 11.34 | 64.71 |
| M | ANN | 20 | 5 | 4 | 9 | 1 | 9.72 | 83.33 |
| | SVM | 20 | — | 4 | 9 | 1 | 8.39 | 66.67 |
| | ANN | 10 | 8 | 5 | 8 | 2 | 9.39 | 82.61 |
| | SVM | 10 | — | 5 | 8 | 2 | 6.69 | 60.87 |
| | ANN | 10 | 6 | 5 | 5 | 3 | 11.04 | 69.23 |
| | SVM | 10 | — | 5 | 5 | 3 | 7.81 | 57.69 |
| I | ANN | 10 | 6 | 5 | 5 | 3 | 14.31 | 61.54 |
| | SVM | 10 | — | 5 | 5 | 3 | 13.55 | 38.46 |
| | ANN | 15 | 6 | 9 | 9 | 3 | 16.72 | 64.71 |
| | SVM | 15 | — | 9 | 9 | 3 | 16.09 | 41.18 |

a small-scaled sample set such as ours. Nonetheless, considering the forecast for volatility trend, the ANN-based approach considerably outplays the other.

### Disturbance experiments

We further justify the correlation between on-line financial information volume and the financial trading volume volatility by conducting disturbance experiments on the well-trained ANN and SVM during the testing process. To accomplish this, we consecutively adjust each element of the input vector $[\sigma^2_{t-1}, \ldots \sigma^2_{t-p}, y^2_{t-1}, \ldots y^2_{t-q}, W^2_{t-1}, \ldots W^2_{t-r}]$ from 75% to 125% of the original value, by a predetermined step-length, and thereafter observe and record the corresponding changing rate of the output incurred. Let $\Delta_t^{i,j}$ denote the changing rate of the forecast volatility on day $t$ caused by adjusting the value of the $i$th element in the input vector by $j$ step-lengths, $\overline{\Delta^{i,j}}$ the average changing rate of the output caused by the above action within the day $t_0$ and the day $t_1$, and we have

$$\overline{\Delta^{i,j}} = \frac{1}{t_1 - t_0 + 1} \sum_{t=t_0}^{t_1} \Delta_t^{i,j}, \tag{27}$$

where

$$\Delta_t^{i,j} = \frac{\overset{\wedge}{\sigma}{}^{2}_{t}{}^{i,j} - \overset{\wedge}{\sigma}{}^{2}_{t}}{\overset{\wedge}{\sigma}{}^{2}_{t}}. \tag{28}$$

The $\overset{\wedge}{\sigma}{}^{2}_{t}{}^{i,j}$ in Eq. (22) represents the changed forecast volatility when the $i$th element of the input vector has been modified by $j$ step-lengths.

In Table 5 is demonstrated the computed values of $\overline{\Delta^{i,j}}$ when conducting the ANN-based disturbance experiment on the index NASDAQ within the period from June 30, 2006 to September 28, 2006. The specific configurations regarding the parameters are $p = 3$, $q = 6$, $r = 4$, $C = 12$, and $H = 7$. The rows in grey indicate the changing rates of the input elements.

The values of $\overline{\Delta^{i,j}}$ we have acquired when conducting the SVM-based disturbance experiment on the same index within the same period of time with the configurations $p = 3$, $q = 6$, $r = 4$, and $C = 12$ are demonstrated in Table 6. The rows in grey indicate the changing rates of the input elements.

Results shown in Tables 5 and 6 clearly demonstrate that on-line financial information (indicated by $W^2_{t-k}$) volume does have a certain influence on the output, which can be considered as of the same magnitude of significance as that of the other input factors.

### Discussion

The disturbance experiments conducted on both approaches disclose the fact that on-line financial information does have an influence on the trading volume volatility, which can be considered as of the same magnitude as those caused by volatilities and daily changing rates within the immediate past. Therefore, our work validates the existence of the associations between on-line financial information and trading volume volatility.

**Table 5.** The ANN-based disturbance experiment results for the index NASDAQ (in percentage).

| Changing rate | 75 | 80 | 85 | 90 | 95 | 105 | 110 | 115 | 120 | 125 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2_{t-1}$ | −5.204 | −3.886 | −2.643 | −1.663 | −0.86 | 1.068 | 2.437 | 3.876 | 5.099 | 6.171 |
| $\sigma^2_{t-2}$ | −1.961 | −1.332 | −0.773 | −0.359 | −0.105 | 0.011 | 0.004 | 0.019 | 0.001 | 0.151 |
| $\sigma^2_{t-3}$ | −3.693 | −2.941 | −2.142 | −1.382 | −0.661 | 0.572 | 1.124 | 1.72 | 2.326 | 2.891 |
| $y^2_{t-1}$ | −0.144 | −0.114 | −0.084 | −0.055 | −0.027 | 0.026 | 0.052 | 0.077 | 0.102 | 0.126 |
| $y^2_{t-2}$ | 0.32 | 0.26 | 0.198 | 0.134 | 0.068 | 0.071 | 0.144 | 0.219 | 0.297 | 0.376 |
| $y^2_{t-3}$ | −0.123 | −0.1 | −0.077 | −0.052 | −0.027 | 0.028 | 0.058 | 0.089 | 0.122 | 0.157 |
| $y^2_{t-4}$ | 0.029 | 0.018 | 0.01 | 0.005 | 0.001 | 0.001 | 0.003 | 0.006 | 0.011 | 0.017 |
| $y^2_{t-5}$ | −0.117 | −0.089 | −0.063 | −0.039 | −0.018 | 0.015 | 0.026 | 0.035 | 0.041 | 0.045 |
| $y^2_{t-6}$ | 0.04 | 0.026 | 0.015 | 0.007 | 0.002 | 0.001 | 0.004 | 0.011 | 0.019 | 0.03 |
| $W^2_{t-1}$ | −0.369 | −0.305 | −0.236 | −0.162 | −0.083 | 0.086 | 0.175 | 0.266 | 0.357 | 0.448 |
| $W^2_{t-2}$ | −0.256 | −0.202 | −0.149 | −0.098 | −0.049 | 0.048 | 0.096 | 0.143 | 0.191 | 0.24 |
| $W^2_{t-3}$ | 0.247 | 0.194 | 0.143 | 0.094 | 0.046 | 0.045 | 0.089 | 0.132 | 0.174 | 0.216 |
| $W^2_{t-4}$ | −0.009 | −0.008 | −0.007 | −0.005 | −0.003 | 0.004 | 0.008 | 0.013 | 0.018 | 0.024 |

We have also applied our models upon the stock price return time series. Empirical studies show that if we take online information volume as an exogenous input, the forecasting performance for the trading volume volatility considerably outplays that for the price return volatility. Therefore, trading volume is more inclined to be affected by on-line financial information. In addition, we have found out that the larger the value of D, the smaller the average forecast error becomes, which further proves the volatility clustering feature of financial time series. Additionally, better forecast performance can be achieved if we square the moving average part of the input vector, which substantiates one of the GARCH theory's conclusions that there is a significant correlation between the squared residuals of financial time series.

**Table 6.** The SVM-based disturbance experiment results (in percentage).

| Changing rate | 75 | 80 | 85 | 90 | 95 | 105 | 110 | 115 | 120 | 125 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2_{t-1}$ | −4.033 | −3.486 | −2.777 | −1.931 | −0.988 | 0.977 | 1.892 | 2.706 | 3.395 | 3.95 |
| $\sigma^2_{t-2}$ | −1.064 | −0.827 | −0.579 | −0.346 | −0.15 | 0.105 | 0.176 | 0.233 | 0.294 | 0.371 |
| $\sigma^2_{t-3}$ | −2.224 | −1.865 | −1.447 | −0.984 | −0.496 | 0.488 | 0.957 | 1.396 | 1.801 | 2.169 |
| $y^2_{t-1}$ | −0.168 | −0.134 | −0.1 | −0.066 | −0.033 | 0.032 | 0.064 | 0.096 | 0.127 | 0.157 |
| $y^2_{t-2}$ | 0.158 | 0.125 | 0.092 | 0.06 | 0.03 | 0.029 | 0.056 | 0.082 | 0.107 | 0.131 |
| $y^2_{t-3}$ | −0.082 | −0.068 | −0.052 | −0.036 | −0.018 | 0.019 | 0.04 | 0.061 | 0.082 | 0.105 |
| $y^2_{t-4}$ | −0.005 | −0.006 | −0.006 | −0.005 | −0.003 | 0.004 | 0.008 | 0.014 | 0.02 | 0.027 |
| $y^2_{t-5}$ | −0.036 | −0.032 | −0.026 | −0.019 | −0.01 | 0.012 | 0.025 | 0.041 | 0.057 | 0.076 |
| $y^2_{t-6}$ | −0.057 | −0.047 | −0.036 | −0.024 | −0.012 | 0.013 | 0.026 | 0.039 | 0.053 | 0.067 |
| $W^2_{t-1}$ | −0.225 | −0.183 | −0.139 | −0.094 | −0.047 | 0.048 | 0.096 | 0.145 | 0.193 | 0.24 |
| $W^2_{t-2}$ | −0.223 | −0.179 | −0.135 | −0.09 | −0.045 | 0.046 | 0.091 | 0.137 | 0.183 | 0.229 |
| $W^2_{t-3}$ | −0.021 | −0.018 | −0.015 | −0.011 | −0.006 | 0.006 | 0.014 | 0.021 | 0.03 | 0.039 |
| $W^2_{t-4}$ | −0.161 | −0.131 | −0.1 | −0.068 | −0.034 | 0.035 | 0.071 | 0.108 | 0.146 | 0.185 |

## Volatility Forecasting with Sentiment Analysis

### Experiment design and results

Correlations between financial news sentiment value and stock price volatility of listed company are investigated using the GARCH-based SVM regression in this section. A time window with a length of 7 days, namely, 5 trading days, is chosen. The forecast is implemented in a progressive fashion, with sliding time windows, and eventually aggregated statistics upon all the companies are obtained. Before fed into the SVM model, both quote and news data are segmented into time windows. The financial news in this experiment is downloaded from more than 200 English portal websites on the Internet.

The SVM toolbox utilized in this experiment is the open source library LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The news entries have a time horizon starting from January 1, 2007 to December 3, 2007, which spans across 49 time windows. In total, a set of 177 listed companies in the U.S. stock markets are studied. There are 153,468 pieces of news for all the companies, all of which are tagged with sentiment values computed using the algorithm in the section titled *Sentiment Analysis for Financial News*. The kernel function we used is the RBF function. Two major performance metrics are introduced in this experiment to evaluate the aggregated forecast performance for all 177 companies: squared correlation coefficient (SCC) and volatility trend forecast accuracy (VTFA). SCC and VTFA are computed based on the forecasting values for each time window.

The ***squared correlation coefficient*** evaluates the correlation of all the explanatory variables to the response variable. The closer this value is to 1, the better regression result is achieved. Squared correlation coefficient is calculated as follows. Suppose there are $M$ listed companies; for a specific time window $W_i$, the actual values of price volatility for company $k$ are denoted as $\sigma_i^{p2}(k)$, where $k = 1,2,\ldots M$. The predicted volatility for company $k$ is therefore denoted as $\hat{\sigma_i^{p2}}(k)$. The average actual value of price volatility for all companies can thus be formulated as

$$\bar{\sigma_i^{p2}} = \frac{1}{M}\sum_{k=1}^{M}\sigma_i^{p2}(k). \tag{29}$$

The $\sigma_i^{p2}(k) - \bar{\sigma_i^{p2}}$ is named as the deviation of the company $k$. Let $S$ denote the deviation of all the company set, thus $S$ is calculated as

$$S = \sum_{k=1}^{M}\left(\sigma_i^{p2}(k) - \bar{\sigma_i^{p2}}\right)^2 = \sum_{k=1}^{M}\left(\sigma_i^{p2}(k) - \hat{\sigma_i^{p2}}(k)\right)^2 + \sum_{k=1}^{M}\left(\hat{\sigma_i^{p2}}(k) - \bar{\sigma_i^{p2}}\right)^2, \tag{30}$$

where

$$U = \sum_{k=1}^{M}\left(\hat{\sigma_i^{p2}}(k) - \bar{\sigma_i^{p2}}\right)^2 \tag{31}$$

reflects the oscillation of the response variable caused specifically by those explanatory variables, and

$$Q = \sum_{k=1}^{M} \left( \sigma_i^{p2}(k) - \hat{\sigma_i^{p2}}(k) \right)^2 \tag{32}$$

denotes the squared sum of residuals, which is an indicator of the system error, usually caused by objective factors.

Therefore the proportion of $U$ of $S$ is an effective indicator of the correlation performance. A big percentage $U$ takes up in $S$, which renders naturally a small percentage $Q$ takes up in $S$, guaranteeing a significant regression performance. Let $R^2$ denote the proportion of $U$ to $S$, thus we have

$$R^2 = \frac{U}{S} = \frac{S - Q}{S}, \tag{33}$$

where the closer the value of $R^2$ is to 1, the better correlation effect between the sentiment and the volatility is proved.

The ***Volatility trend forecast accuracy*** is a proportion of the companies with the accurately predicted volatility trend among all the companies. The definition of volatility trend is consistent with the first experiment. In Table 7 is presented a demonstration of the asset price volatility forecast for 177 companies using information sentiment during year 2007, from the third week till the fiftieth week. Considering space limitations, only part of the results is shown here. Note that the penalty parameter c and the RBF kernel parameter g is set as c = 64 and g = 1/3.

It can be observed that an average of 60.3225% is achieved for the volatility trend forecast and an average of 71.2593% is achieved for the squared correlation coefficient upon all 177 companies with the 48 forecast time windows. In Figures 10 and 11 are visualized the price volatility forecasts for two specific companies out of the 177, which are MDT and WAG. In Figure 12 are shown the VTFA for all the time windows, corresponding to Table 7.

**Table 7.** Par of the forecasting results for 177 listed companies within the year 2007.

| Time window for forecast output | Time window for forecast input | Time windows for training samples | SCC ($R^2$) | VTFA |
|---|---|---|---|---|
| Week3 | Week2 | Week1–2 | 0.662534 | 0.655367 |
| Week4 | Week3 | Week2–3 | 0.669676 | 0.525424 |
| Week5 | Week4 | Week3–4 | 0.873588 | 0.59887 |
| Week6 | Week5 | Week4–5 | 0.248732 | 0.694915 |
| Week7 | Week6 | Week5–6 | 0.584079 | 0.59887 |
| Week8 | Week7 | Week6–7 | 0.89559 | 0.632768 |
| Week9 | Week8 | Week7–8 | 0.987398 | 0.305085 |
| Week10 | Week9 | Week8–9 | 0.98691 | 0.649718 |
| ... | ... | ... | ... | ... |
| Week48 | Week47 | Week46–47 | 0.928572 | 0.508475 |
| Week49 | Week48 | Week47–48 | 0.724147 | 0.672316 |
| Week50 | Week49 | Week48–49 | 0.999941 | 0.553672 |

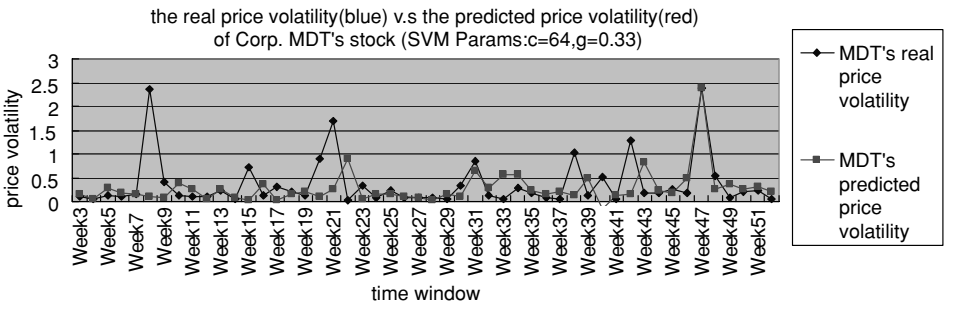


**Figure 10.** Price volatility forecast result for company MDT over all the time windows.



## DISCUSSION

Experiments have been performed on both trading volume volatilities and asset price volatilities, whereas in contrast to the conclusion in the first experiment, it is proved that asset price bears a closer correlation with information sentiment, considering the fact that a better forecast performance is achieved upon the price volatility. As shown in Figure 9 and Figure 10, the predicted values, under most circumstances, correspond to the actual ones well, whereas for an occasional huge oscillation, the forecast result is not very satisfying.

For both asset price volatility and trading volume volatility forecasts, an average of more than 60% of VTFA can be achieved, which substantiates the existence of firm correlations between information sentiment and volatility trend. It is also discovered by the experiment that during the days when there is a large number of news, incorporating information sentiment into the machine learning model is able to noticeably increase the volatility trend forecasting. Besides, an average of more than 70% of SCC was achieved for both price volatility and trading volume volatility forecasting, which in addition gives convincing proof to the correlations in between.

What makes the experiment in this step different from the first is that it extends its study to a large set of companies. Empirical results are reflected by aggregated

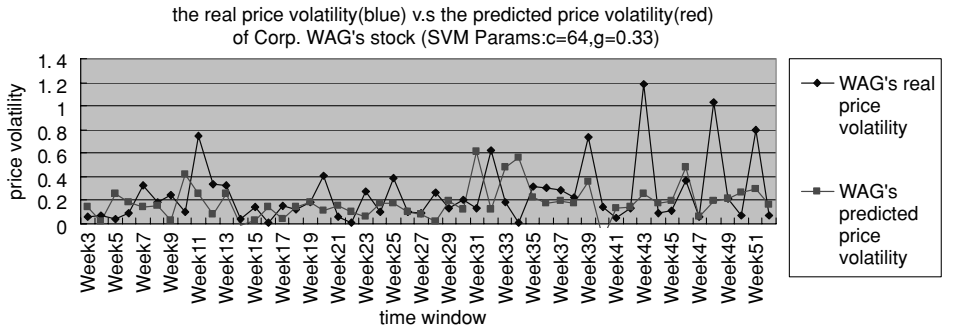


**Figure 11.** Price volatility forecast result for company WAG over all the time windows.

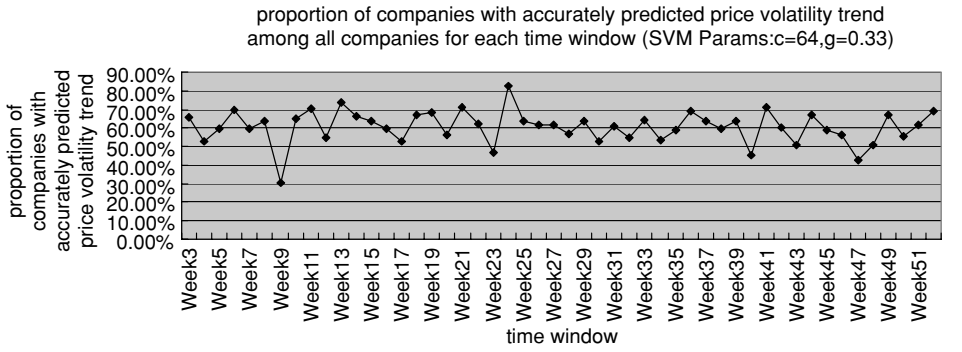


**Figure 12.** Price volatility trend forecast accuracies for all the time windows.



statistics, which indicates the effects of information unto the whole of the stock markets. The results of this phase, although only focused on U.S. markets currently, provide vivid description of the macro influences of financial news on financial volatility. It is of critical value to financial practitioners, who seek to get a panorama picture of the general market, in strategic decision-making.

## CONCLUSION AND FUTURE WORK

We have introduced in this article the GARCH-based ANN and SVM to investigate the correlations between financial trading volume volatility and on-line information volume and thus effectively predict financial risk under network environment. Both methods are capable of achieving favorable predicting results. The GARCH-based ANN performs better in predicting the volatility trend than the GARCH-based SVM, whereas the GARCH-based SVM outperforms the GARCH-based ANN in forecasting the volatility itself. Moreover, on-line information is converted to sentiment values, which constitutes another key input element for the machine learning models. Empirical studies indicate solid correlations between asset price volatility and information sentiment, which is well captured and stored by the SVM. Aggregated statistics show that a good forecast performance can be achieved by use of the GARCH-based SVM method under sliding time windows. These empirical studies are advisory and helpful for financial investors, portfolio holders, academicians, and so on in the sense that they provide an alternative tool to forecast the volatility and the trend.

Future work is still necessary. The first direction is to expand the dataset we used to generate more useful results. The second important direction is to develop a more efficient sentiment calculation algorithm in order to enhance the accuracy of judging the emotional polarity of articles.

## ACKNOWLEDGMENT

The 863 Project of China under grant number 2007AA01Z437 is acknowledged.

## REFERENCES

Ahmad K and Almas Y. 2005. Visualising sentiments in financial texts? Proceedings of the Ninth International Conference on Information Visualisation 1:363–8. London, UK

Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. J Econometrics 31(3):307–27

Burges C. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2):121–67

Cai CL and Shi ZZ. 2003. A modular neural network architecture with approximation capability and its applications. Proceedings of the Second IEEE International Conference on Cognitive Informatics, pp 60–4. London, UK

Chaovalit P and Zhou L. 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. Proceedings of the 38th Hawaii International Conference on System Sciences. Big Island, HI, USA

Chuttur MY and Bhurtun C. 2005. Monitoring financial market using French written textual data. IEEE 3rd International Conference on Computational Cybernetics, pp 239–42. Budapest, Hungary

Costantino M, Morgan RG, Collingham RJ, *et al.* 1997. Natural language processing and information extraction: Qualitative analysis of financial articles. Proceedings of the IEEE/IAFE, pp 116–22. New York, USA

Cristianini N and Shawe-Taylor J. 2000. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK

Dong ZD and Dong Q. 2003. HowNet—A hybrid language and knowledge resource. Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering, pp 820–4. Beijing, China

Eitrich T and Lang B. 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. J Computational App Math 196(2):425–36

Freisleben B and Ripper K. 1997. Volatility estimation with a neural network. Proceedings of the IEEE/IAFE on Computational Intelligence for Financial Engineering, pp 177–81. New York, USA

Hu JL, Wu Z, McCann H, *et al.* 2006. BFGS quasi-Newton method for solving electromagnetic inverse problems. Microwaves, Antennas and Propagation. IEEE Proceedings 153(2):199–204

Huang GB, Chen L, and Siew CK. 2006. Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Transactions on Neural Networks 17(4):879–92

Kanzow C, Yamashita N, and Fukushima M. 2004. Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints. J Computational App Math 172(2):375–97

Sohn SY and Lim M. 2007. Hierarchical forecasting based on AR-GARCH model in a coherent structure. European J Operational Res 176(2):1033–40

Tino P, Schittenkopf C, and Dorffner G. 2001. Financial volatility trading using recurrent neural networks. IEEE Transactions on Neural Networks 12(4):865–74

Tong S and Chang E. 2001. Support vector machine active learning for image retrieval. Proceedings of ACM International Conference on Multimedia, pp 107–18. Ottawa, On, Canada

Trafalis TB and Ince H. 2000. Support vector machine for regression and applications to financial forecasting. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks 6:348–53. Como, Italy

Turney PD. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning, pp 491–502. Springer-Verlag, Berlin, Germany

Turney PD. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, NJ, USA

Turney PD and Littman ML. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21:315–46

Vapnik V. 1998. Statistical Learning Theory. John Wiley, New York, USA

Wang HF and Hu DJ. 2005. Comparison of SVM and LS-SVM for regression. International Conference on Neural Networks and Brain 2005 1:279–283. Beijing, China

Wu CH, Ho JM, and Lee DT. 2004. Travel-time prediction with support vector regression. IEEE Transactions on Intelligent Transportation System 5(4):276–81

Wu D, Liang L, and Yang Z. 2007. An empirical study of financial distress of Chinese public companies using PNN and MDA. Socio-Economic Planning Sciences (*in press*)

Zhang XH, Lu ZB, and Kang CY. 2003. Underwater acoustic targets classification using support vector machine. Proceedings of the International Conference on Neural Networks and Signal Processing 2:932–5. Houston, TX, USA