Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2012 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

7-15-2012

Web Mining For Financial Market Prediction Based On Online Sentiments

Wei Xu
School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, weixu@ruc.edu.cn

Tuo Li
School of Information, Renmin University of China, Beijing, P.R. China, lituo1990@gmail.com

Bing Jiang
School of Information, Renmin University of China, Beijing, P.R. China, jiangbingruc@163.com

Cheng Cheng
School of Information, Renmin University of China, Beijing, P.R. China, chengcheng_ruc@126.com

WEB MINING FOR FINANCIAL MARKET PREDICTION BASED ON ONLINE SENTIMENTS

Wei Xu, School of Information, Renmin University of China, and Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, Beijing, P.R. China, weixu@ruc.edu.cn

Tuo Li, School of Information, Renmin University of China, Beijing, P.R. China, lituo1990@gmail.com

Bing Jiang, School of Information, Renmin University of China, Beijing, P.R. China, jiangbingruc@163.com

Cheng Cheng, School of Information, Renmin University of China, Beijing, P.R. China, chengcheng_ruc@126.com

Abstract

Financial market prediction is a critically important research topic in financial data mining because of its potential commerce application and attractive profits. Previous studies in financial market prediction mainly focus on financial and economic indicators. Web information, as an information repository, has been used in customer relationship management and recommendation, but it is rarely considered to be useful in financial market prediction. In this paper, a combined web mining and sentiment analysis method is proposed to forecast financial markets using web information. In the proposed method, a spider is firstly employed to crawl tweets from Twitter. Secondly, Opinion Finder is offered to mining the online sentiments hidden in tweets. Thirdly, some new sentiment indicators are suggested and a stochastic time effective function (STEF) is introduced to integrate everyday sentiments. Fourthly, support vector regressions (SVRs) are used to model the relationship between online sentiments and financial market prices. Finally, the selective model can be serviced for financial market prediction. To validate the proposed method, Standard and Poor's 500 Index (S&P 500) is used for evaluation. The empirical results show that our proposed forecasting method outperforms the traditional forecasting methods, and meanwhile, the proposed method can also capture individual behavior in financial market quickly and easily. These findings imply that the proposed method is a promising approach for financial market prediction.

Keywords: Web mining, sentiment analysis, statistical learning theory, financial market, prediction

1 INTRODUCTION

Financial market prediction is one of the most significant research areas in financial data mining because of its commerce application and kinds of attractive profits (Lee, 2004). However, it is also regarded as a challenging task for its high volatility, irregularity, and noise in financial market (Yu et al., 2009). Although financial market prediction is extremely difficult, due to research challenge and benefit inspiration, many scholars and practitioners pay much attention. Accordingly, hundreds of models and methods have been proposed to forecast financial market. Initially, time-series forecasting methods have been suggested for financial market prediction by various techniques (Dacorogna et al., 1996; Wang, 2002; Tseng et al., 2002; Hung, 2009; Bildirici and Ersin, 2009; Liao and Wang, 2010; Xu et al. 2011). For example, Dacorogna et al. (1996) proposed a short-term forecasting model in financial market by introducing the concept of intrinsic time. Xu et al. (2011) offered a neural network based ensemble method for financial time series prediction. Wang (2002) used a fuzzy grey method for predicting stock prices, while Liao and Wang (2010) developed a stochastic time effective neural network. Hung (2009) employed a fuzzy GARCH approach to model the price index. Furthermore, Tseng et al. (2002) combined neural network with ARIMA model. Similarly, Bildirici and Ersin (2009) hybridized GARCH family models with the artificial neural networks for Istanbul stock exchange prediction.

Technical analysis, as an important analysis tool in financial market, has been introduced in financial market prediction. For example, Refenes and Holt (2001) used technical indicators to forecast volatility with neural regression. Rodriguez-Gonzalez et al (2010) offered a RSI indicator to improve trading systems using neural network. Nair et al. (2010) suggested a decision tree-rough set hybrid system for stock market trend prediction using technical indicators. Similarly, Hsu (2011) selected16 technical indicators to forecast stock market by using the SOM-GP procedure. Also, economic/financial variables are suggested to analyze financial market. Kanas and Yannopoulos (2001) compared linear and nonlinear predictors in stock market using financial indicators. Similarly, Olson and Mossman (2003) suggested using various accounting ratios to forecast Canadian stock returns. Cao et al. (2005) compared Fama and French's model and artificial neural networks in predicting stock market. Chiu and Chen (2009) combined technical indicators and macroeconomic variables for exploring internal mechanism of stock market using fuzzy based support vector machines.

Recently, web information, as an information repository, has been added to model the financial market trend. Mittermayer (2004) analyzed intraday stock price trends based on text mining techniques, and the findings indicate that press releases is able to provide additional information that can be used for mining financial market trend. Schumaker and Chen (2009) examined the problem of financial market prediction by a synthesis of linguistic, financial and statistical techniques, and have been applied to financial market. Similarly, Schumaker and Chen (2010) developed a discrete stock price prediction engine based on financial news, and experimental results demonstrate that the proposed engine outperforms the market average and performs well against existing quant funds. Furthermore, Li et al. (2010) forecasted financial market prices based on news density. Chan and Franklin (2011) constructed a text-based decision support system for financial sequence prediction, and empirical results show that the proposed system outperforms similar statistical models based on prediction accuracy.

Different from previous studies, Das and Chen (2007) extracted sentiment from small talks on the web, and examined the relationship between sentiment and the stock values.In their methodology, the proposed algorithms can be applied to assess the impact on investor opinion of management announcements, regulatory changes, and press releases. Sehgal and Song (2007) introduced a new method to forecast stock market based on online sentiments. In their proposed method, financial message boards were scanned, online expressed sentiments were extracted, and the relationship of stock market and sentiments are validated by using trust calculation. Furthermore, Bollen and Mao (2011) offered twitter mood as a stock market predictor to model the relationship between online

emotions and stock prices, and the experimental results indicate that online emotions can play an important role in financial markets. Similarly, Bollen et al. (2011) analyzed how to extract twitter mood from different dimensions, and how to combine online emotions for stock market prediction in details.

As mentioned in the literature (Bollen et al, 2011), behavioural finance shows us that emotions may profoundly affect individual behaviour and decision-making, and the scholars in behavioural finance research now can apply computational models to large-scale social media data for better explaining and forecasting financial market. Although the emotions may have a great influence on the investors' behaviour, however, the degree of impact of the motions depends on the time at which they occurred (Liao and Wang, 2010). Also, different dimensions of emotions (positive or negative) can affect investors' behaviour differently, some dimensions may have greater influence, and some may not. In this paper, a new online social sentiment based method is proposed for financial market prediction using web mining. In our proposed method, a spider is firstly used to crawl tweets from Twitter. Secondly, OpinionFinder is utilized to mining the online sentiments hidden in tweets. Thirdly, several novel sentiment indicators are suggested and a stochastic time effective function (STEF) is introduced to integrate everyday sentiments. After all these, support vector regressions (SVRs) are used to model the association between online sentiments and financial market prices. Finally, the selective model can be serviced for financial market prediction. In the whole process, the parameters in the proposed method are discussed in detail, and also the proposed method is validated by real financial market data.

The rest of this paper is organized as follows. Some theories including a stochastic time effective function and statistical learning theory, which are used in our proposed method, are presented in Section 2. Section 3 provides the details for our proposed online sentiment based method. For the purpose of validation, empirical analysis of the proposed method in real financial market prediction is reported in Section 4.Finally, conclusions and future research directions are summarized in Section 5.

2 THEORETICAL BACKGROUND

In this section, some useful theories used in our proposed method are briefly introduced. In detail, the basic concepts of a stochastic time effective function are described, and the statistical learning theory is also summarized. These theories are effective and efficient to develop our proposed method.

2.1 Introduction to A Stochastic Time Effective Function

The stochastic time effective function (STEF), proposed by Liao and Wang (2010), has been used to forecast the index of financial market. The general idea of a stochastic time effective function is that the information from different time has different effects. In detail, the more recent the information is, the more effect the information will have. The more recent information will have a higher weight than that of previous one and play a more important role. The widely use of Exponential Moving Average (EMA) in technical analysis of stock price supports that the general idea of STEF is applicable in stock price prediction. The stochastic time effective function is defined as follow:

$$\phi(t_1 \sim t_n) = \frac{1}{\tau}\exp\{\int_{t_1}^{t_n} \mu(t)dt + \int_{t_1}^{t_n} \sigma(t)dB(t)\} \quad (1)$$

where τ ($\tau > 0$) stands for the time strength coefficient, t_1 means the most recent time or the time where relative data with the highest weight, and t_n defines an arbitrary time in dataset. $\mu(t)$, $\sigma(t)$ and $B(t)$ are the drift function, volatility function, and standard Brownian motion respectively.

2.2 Introduction to Statistical Learning Theory

Statistical learning theory, proposed by Vapnik (1995), is firstly used for classification. After the introduction of Vapnik's ε-insensitivity loss function, a support vector regression (SVR), which is

especially designed for solving nonlinear estimation problems, is generated. After that, SVR has been successfully applied for the prediction of financial time series (Lu et al., 2009). This subsection will provide a brief introduction of basic concept of SVR, especially ε-SVR.

In typical regression problem, given a training set $\{(x_i, y_i)\}_{i=1}^{N}$, the target of SVR is to simulate a regression function $y = f(x)$ such that the trained SVR model will accurately predict the value of y_i given a new input x_i.

Generally speaking, the purpose of ε-SVR is to find $f(x)$ that will place most of the data in the tube, as shown in Figure 1. Parameter ε stands for precision parameter which representing the radius of the tube and the region enclosed by the tube is named ε-insensitive zone.

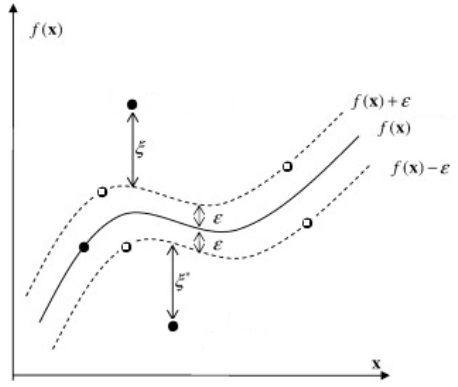

Figure 1. An illustration of ε-SVR

In SVR, the regression is defined as:

$$f(x) = (v \cdot \phi(x)) + b \tag{2}$$

where v denotes weight vector and b is constant, $\phi(x)$ denotes a mapping function from input space to high dimension feature space. Here in SVR, data are firstly mapped from input space to high dimension feature space by $\phi(x)$ and then a linearly regression is carried out. The coefficients of regression are estimated by minimizing the regularized risk function:

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(y_i, f(x_i)) + \frac{1}{2}\|w\|^2 \tag{3}$$

where C is the regularization constant which is used to strike the balance between training error and flatness of model, and $\frac{1}{2}\|w\|^2$ is regularization term which is used to control the balance between complexity and accuracy of SVR. $L_{\varepsilon}(y, f(x))$ is the ε-insensitivity loss function, which is defined as:

$$L_{\varepsilon}(y, f(x)) = \begin{Bmatrix} 0 & if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{Bmatrix} \tag{4}$$

When slack variables ξ are introduced, the SVR formulation can be transformed into convex optimization problem:

$$
\begin{aligned}
\textit{Minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_j^*) \\
& (y_i - (w\cdot\phi(x)+b)) \le \varepsilon + \xi_i, \\
\textit{s.t.} \quad & (y_j - (w\cdot\phi(x)+b)) \ge \varepsilon + \xi_j^* \\
& \xi_i, \xi_j^* \ge 0
\end{aligned} \tag{5}
$$

By introducing Lagrange multipliers and exploiting optimality constraints, the formula can be transformed into its dual problem:

$$
f(x,\alpha,\alpha^*) = \sum_{i=1}^{Nsv}(\alpha_i - \alpha_j^*)K(x,x_i) + b \tag{6}
$$

where $K(x,x_i)$ is called kernel function and $K(x,x_i) = \phi(x)\cdot\phi(x_i)$. α and α^* are the Lagrange multipliers.

The kernel function plays an important role in mapping input vector into a high dimensional space. Linear kernel, polynomial kernel, radial basis function kernel and sigmoid kernel are commonly used kernels in SVM. Among these kernels, radial basis function kernel is widely used in financial time series prediction (Chi-Jie Lu et al., 2009). The Radial Basis Function (RBF) kernel is represented as follows:

$$
K\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), \gamma > 0 \tag{7}
$$

3 THE PROPOSED FORECASTING METHOD

In this section, an online social sentiment based method is proposed for financial market prediction.In our proposed model, a spider is firstly employed in crawling tweets from Twitter. Then, OpinionFinder is utilized to mining the online social sentiments hidden in tweet; After that, two new sentiment indicators are proposed, and besides, a stochastic time effective function (STEF) is introduced to integrate the mined sentiments. Next, support vector regressions (SVRs) are employed to model the connection between online social sentiments and financial market prices. Finally, the selective model can be utilized for financial market prediction. The framework of the proposed method is illustrated in Figure 2, and the details will be discussed in the following subsections.

3.1 Crawling Tweets from Twitter

The tweet dataset from the Stanford Network Analysis Project (SNAP)(http://snap.stanford.edu/data/twitter7.html)is selected instead of crawling ourselves, because of the inaccessibility of Twitter in China. The whole dataset, which covers about 20%-30% of all public tweets published on Twitter from June 1st 2009 to December 31st 2009, contains about 467 million Twitter posts from 20 million users. The tweets, which containing the basic information of the submission time, identity number of the author and comments, are suitable for us to extract sentiments of the public. The dataset is firstly examined by us manually and then the unusual data are eliminated. Finally, the tweets from August to December are taken into consideration for our experiments. What's more, the historical stock prices of S&P 500 Index are picked up within the same time interval from Yahoo Finance (http://finance.yahoo.com).

3.2 Mining online social sentiment

During the process of valuing the sentiments of tweets, the MPQA (Multi-Perspective Question Answering) Subjectivity lexicon (http://www.cs.pitt.edu/mpqa/subj_lexicon.html) is adopted, and

OpinionFinder is employed to classify the strength and polarity of a single word's sentiment by using MPQA. The most frequently used 6443 English words, in terms of their strength and polarity, are classified into five degrees which are strong positive, weak positive, neutral, weak negative and strong negative respectively. In this way, the strength and polarity of every single tweet can be calculated. Thus, through this classification, it can not only calculate the positive or negative values of the words in tweets, but also make a further compelling exploration on the prediction of the strong or weak emotion in tweets.

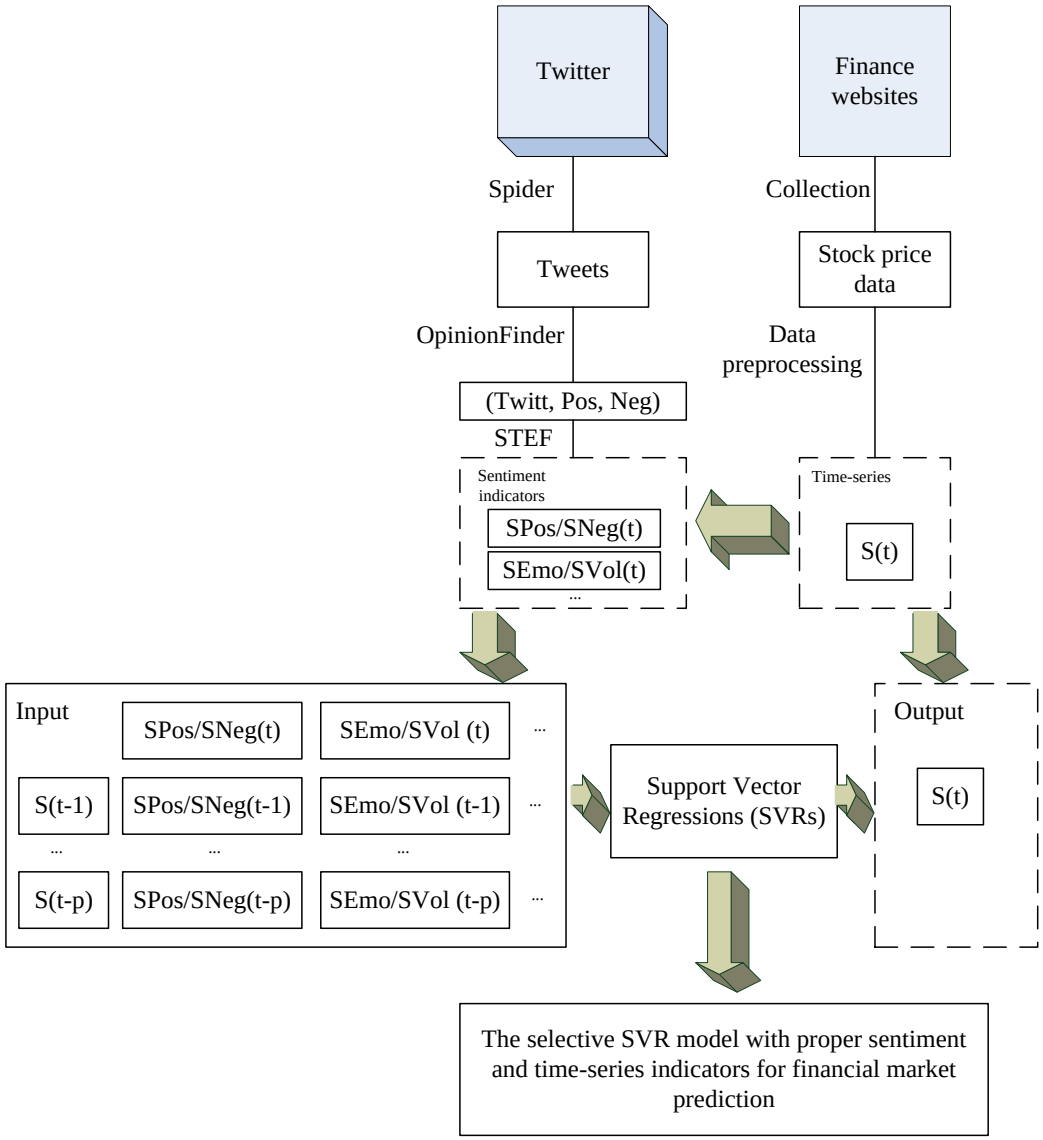

Figure 2. The framework of the proposed method

3.3 Developing sentiment indicators

Preliminarily, some processes and considerations are made to extract the public sentiments from tweets. Firstly, the number of tweets on a certain time interval (the interval length is one hour in this paper) is courted to reflect the volume of information as well as activeness of the users. Next, the sentiments such as the overall positive and negative sentiments are calculated, and also the integral emotion, which is the difference of the two sentiments above, is computed. Finally, a stochastic time effective function aforementioned is introduced to recalculate the indicators in a certain time window in adapted way. Formulations and specifics are listed below.

Indicators without incorporating stochastic time effective function are listed below:

$$Pos_{(t)} = \sum_{n=1}^{Vol_{(t)}} pos_{t_n} \tag{8}$$

$$\mathrm{Neg_{(t)}} = \sum_{\mathrm{n=1}}^{\mathrm{Vol_{(t)}}} \mathrm{neg_{t_n}} \tag{9}$$

$$\mathrm{Emo_{(t)}} = \sum_{\mathrm{n=1}}^{\mathrm{Vol_{(t)}}} (\mathrm{pos_{t_n}} - \mathrm{neg_{t_n}}) \tag{10}$$

Indicators that incorporated with stochastic time effective function are listed below:

$$\mathrm{SPos_{(t)}} = \sum_{\mathrm{n=1}}^{\mathrm{Vol_{(t)}}} \phi(\mathrm{t_1 \sim t_n})\mathrm{pos_{t_n}} \tag{11}$$

$$\mathrm{SNeg_{(t)}} = \sum_{\mathrm{n=1}}^{\mathrm{Vol_{(t)}}} \phi(\mathrm{t_1 \sim t_n})\mathrm{neg_{t_n}} \tag{12}$$

$$\mathrm{SEmo_{(t)}} = \sum_{\mathrm{n=1}}^{\mathrm{Vol_{(t)}}} \phi(\mathrm{t_1 \sim t_n})(\mathrm{pos_{t_n}} - \mathrm{neg_{t_n}}) \tag{13}$$

where $\mathrm{Vol_{(t)}}$ is the number of total tweets on a time interval t, $\mathrm{t_1}$ is the close quotation time on day t, and $\mathrm{t_n}$ is the arbitrary time. $\mathrm{pos_{t_n}}$ and $\mathrm{neg_{t_n}}$ are the positive and negative sentiments of tweets in time interval $\mathrm{t_n}$ respectively. $\mathrm{Pos_{(t)}}$ and $\mathrm{Neg_{(t)}}$ stand for the positive and negative sentiment indicators of tweets on a given time window t.$\mathrm{Emo_{(t)}}$ is the integral emotion indicator. The method for nomination goes the same in STEF situations.

Previous studies in stock price prediction using online sentiments mainly focus on the quantity or score of positive and negative sentiments. However, the strength of sentiments is often ignored. Meanwhile, neither the volume nor the integral emotion could be separately used to review sentiments directly. By combining them together to obtain the emotion density, the problems would be mitigated. So, in this paper, two new sentiment indicators including sentiment ratio and sentiment density are incorporated to test their justification. Formulations and specifics about indicators are listed below.

$$\frac{\mathrm{Pos}}{\mathrm{Neg}}(\mathrm{t}) = \frac{\mathrm{Pos_{(t)}}}{\mathrm{Neg_{(t)}}} \tag{14}$$

where $\frac{Pos}{Neg}(t)$ stands for the ratio of positive sentiment to negative sentiment.

$$\frac{\mathrm{Emo}}{\mathrm{Vol}}(\mathrm{t}) = \frac{\mathrm{Emo_{(t)}}}{\mathrm{Vol_{(t)}}} \tag{15}$$

where$\frac{Emo}{Vol}(t)$ stands for the sentiment density.

Indicator $\frac{\mathrm{Pos}}{\mathrm{Neg}}$(t) reflects the ratio of positive sentiment indicator to negative sentiment indicator on a given time window t, ranging from 0 to infinite. When $\frac{\mathrm{Pos}}{\mathrm{Neg}}$(t)>1, the overall public sentiment is positive and the larger value of $\frac{\mathrm{Pos}}{\mathrm{Neg}}$(t), the higher intensity of positive sentiments. Instead, when $\frac{\mathrm{Pos}}{\mathrm{Neg}}$(t) <1, the overall public sentiment is negative and the smaller value of it, the higher intensity of negative sentiments. Meanwhile, indicator $\frac{\mathrm{Emo}}{\mathrm{Vol}}$(t) reflects the overall public sentiment intensity on a given time window t, ranging from 0 to infinite either. It is reasonable to assume that the larger value of $\frac{\mathrm{Emo}}{\mathrm{Vol}}$(t), the stronger intensity of public emotion.

By introducing the STEF, two indicators above can be denoted as follows:

$$\frac{\mathrm{SPos}}{\mathrm{SNeg}}(\mathrm{t}) = \frac{\mathrm{SPos_{(t)}}}{\mathrm{SNeg_{(t)}}} \tag{16}$$

where $\frac{SPos}{SNeg}(t)$ stands for the ratio of positive sentiment to negative sentiment.

$$\frac{\mathrm{SEmo}}{\mathrm{SVol}}(\mathrm{t}) = \frac{\mathrm{SEmo}_{(\mathrm{t})}}{\mathrm{SVol}_{(\mathrm{t})}} \tag{17}$$

where $\frac{SEmo}{SVol}(t)$ represents the sentiment density.

Both indicator $\frac{\mathrm{SPos}}{\mathrm{SNeg}}$(t) and indicator $\frac{\mathrm{SEmo}}{\mathrm{SVol}}$(t) reflect the same meanings of $\frac{\mathrm{Pos}}{\mathrm{Neg}}$(t) and $\frac{\mathrm{Emo}}{\mathrm{SVol}}$(t) respectively. The difference between these two sets of indicators is whether STEF is employed in calculation.

3.4 Modelling the Relationship using SVRs

Generally speaking, baseline models are firstly constructed for comparison. After that, two online public sentiments are involved in the model. STEF is then introduced for comparison and at last, the impact of information from different time window is tested. In detail, the relationship of the online sentiments and stock prices are modelled in four steps and note that closing price on a given trading day t is denoted as S(t) in the following paragraphs.

Step 1: Constructing the baseline model. Traditional time series model using support vector regression is established as a baseline model. It is clear that the stock prices are relevant time series in which the closing prices of prior days would somewhat affect the future one. Consequently, time lagging series of previous days would improve the accuracy of model. In sum, the moving average of the p (p = 1, …, 5) previous days' closing price, i.e. $\frac{1}{p}$(S(t-1)+ …+S(t-p)) is chosen as input of SVR to predict S(t). The model is taken as baseline for the followers.

Step 2: Incorporating online sentiments. After the baseline is constructed, two new sentiment indicators are separately or jointly added into models for improving the performance of the forecasting models. Time lag of both stock prices and sentiment indicators are also considered.

Step 3: Involving STEF. Grounded the statement of a stochastic time effective function, tweets on a same day are not homogeneous due to their different submission time. The weight are reasonably weakened and mitigated as the time going by. Based on the assumption, a tweet would be of higher influence if it is near to the closing time. Thus, STEF comes to work in the model.

Step 4: Investigating the impact of different time windows. In this step, concerns are focused on how to extract the sentiments with less amount data and meanwhile without changing the reliability of the models. This could be economic for exploring the pragmatic use. Indicators are recalculated and experiments are performed during different time windows.

3.5 Forecasting Financial Market Trend

The S&P 500 index is selected as the real world financial market data for training and testing the proposed models. Generally speaking, the S&P 500 index with the same period of time of our sentiments data from twitter would be used to test whether public sentiments would help in the prediction of financial market prices. When the models with different parameters are validated, the model with the appropriate parameters can be selected and then used to forecast the trend of the financial market.

4 EMPIRICAL ANALYSIS

4.1 Experimental Design and Evaluation Criteria

Tweets are limited only to 140 characters and are proper to reflect the individual opinions or feelings on a given time. With an assembly of large quantity of individual feeling can make a generalization about the public sentiment. Most of the tweets are made by American users, which is an advantage to predict the American financial market. Trying to reduce the arduous work and make the model could

be used effectively and economically, two time windows are chosen to calculate the indicators and incorporate into the model, as listed in Table 1.

Time Window	Beginning	Ending
Half-on trading day	9:30 am, $day_{(t)}$	16:00 pm, $day_{(t)}$
Whole trading day	16:00 pm, $day_{(t-1)}$	16:00 pm, $day_{(t)}$

Table1. The time window in our experiments

The half-on trading day is the time period when the stock market is actually opening on trading day, while the whole trading day begins from yesterday's closing quotation. Theoretically, the whole trading day have more information thus could be more reliable than only the half-on trading day is proposed. But if the latter could in reality equal to the former one, it saves great effort and time in practical implement.

Also, some experiments are designed to illuminate effectiveness and efficiency of proposed method. The details can be seen from Table 2.

Models	Input	Time Lag	Output
Baseline	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)})$	p=1,2,…,5	$S_{(t)}$
Model A	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{Pos}{Neg_{(t)}}+\ldots+\frac{Pos}{Neg_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$
Model B	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{Emo}{Vol_{(t)}}+\ldots+\frac{Emo}{Vol_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$
Model C	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{Pos}{Neg_{(t)}}+\ldots+\frac{Pos}{Neg_{(t-p+1)}}), \frac{1}{p}(\frac{Emo}{Vol_{(t)}}+\ldots+\frac{Emo}{Vol_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$
Model D	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{SPos}{SNeg_{(t)}}+\ldots+\frac{SPos}{SNeg_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$
Model E	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{SEmo}{SVol_{(t)}}+\ldots+\frac{SEmo}{SVol_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$
Model F	$\frac{1}{p}(S_{(t-1)}+\ldots+S_{(t-p)}), \frac{1}{p}(\frac{SPos}{SNeg_{(t)}}+\ldots+\frac{SPos}{SNeg_{(t-p+1)}}), \frac{1}{p}(\frac{SEmo}{SVol_{(t)}}+\ldots+\frac{SEmo}{SVol_{(t-p+1)}})$	p=1,2,…,5	$S_{(t)}$

Table2. The framework of the experiment design

Furthermore, the evaluation criteria employed are Root Mean Squared Error (RMSE) and Correlation Coefficient (CC). RMSE is widely used in regression evaluation. However, for the calculating of square of difference, the effect of outliers might be sometimes exaggerated. CC is a statistical way of measuring the correlation between values, where 1 stands for perfect correlation, -1 stands for negatively perfect correlation and 0 stands for no correlation of the values. But because of its scale independency, CC has its own limitations. The calculation of RMSE and CC is shown below:

$$RMSE = \sqrt{\frac{(f_1 - a_1)^2 + \ldots + (f_n - a_n)^2}{n}} \tag{18}$$

$$CC = \frac{\sum_i (f_i - \bar{f})(a_i - \bar{a})}{\sqrt{\sum_i (f_i - \bar{f})^2 \cdot \sum_i (a_i - \bar{a})^2}} \tag{19}$$

where f_i represents the predicted value, and a_i represents the actual value.

4.2 Experimental Results

According to the process described in section 3, the experiments are carried out by WEKA (Waikato Environment for Knowledge Analysis). For validating the efficiency and effectiveness of the proposed method, daily data of S&P index from 1st July 2009 to 31st December 2009 are taken as dataset, and the proposed method is verified by 10-fold cross validation. Two evaluation criteria including RMSE and CC are used to analyze the experimental results. The results of RMSE are listed in Table 3.

Models	Time Window	Time Lag *p*				
		1	2	3	4	5
Baseline	-	37.06	36.85	36.89	36.77	36.76
Model A	Half-on trading day	36.71	36.45	36.50	36.26	36.32
	Whole trading day	36.72	36.45	36.50	36.26	36.32
Model B	Half-on trading day	36.71	36.45	36.50	36.26	36.32
	Whole trading day	36.71	36.45	36.50	36.26	36.32
Model C	Half-on trading day	36.42	36.16	36.21	35.89	35.98
	Whole trading day	36.42	36.16	36.21	35.89	35.98
Model D	Half-on trading day	36.71	36.50	36.50	36.26	36.32
	Whole trading day	36.72	36.45	36.50	36.26	36.32
Model E	Half-on trading day	36.71	36.45	36.50	36.26	36.32
	Whole trading day	36.71	36.45	36.50	36.26	36.32
Model F	Half-on trading day	36.42	36.16	36.21	35.89	35.98
	Whole trading day	36.42	36.16	36.21	35.89	35.98

Table3. The experimental results in terms of RMSE

As can be seen from Table 3, the first thing which is noticeable is that the RMSE of the baseline model ranges from 36.76 to 37.06. And from model A, model B and model C, the statistics reveals that the incorporation of sentiment indicators, no matter jointly or separately, will slightly improve the performance of model, though not significantly. Besides, sentiment indicators work well when they are introduced to model jointly, as it can be seen from comparing model C and model F with model A, model B and model D, model E respectively. Another thing that this table shows is that the introduction of STEF does not affect the performance of models, as it can be seen when comparing model A, model B and model C with model D, model E and model F respectively. It is also noticeable that the time window of information does not influence the performance of models in general, except in model A with time lag p=1 and model D with time lagging p=1 and 2.

Also, the results of CC are listed in Table 4.

Models	Time Window	Time Lag *p*				
		1	2	3	4	5
Baseline	-	0.64	0.65	0.66	0.67	0.53
Model A	Half-on trading day	0.66	0.68	0.70	0.70	0.58
	Whole trading day	0.66	0.68	0.70	0.70	0.58

Model B	Half-on trading day	0.66	0.68	0.70	0.70	0.58
	Whole trading day	0.66	0.68	0.70	0.70	0.58
Model C	Half-on trading day	0.67	0.69	0.71	0.72	0.61
	Whole trading day	0.67	0.69	0.71	0.72	0.61
Model D	Half-on trading day	0.66	0.70	0.70	0.70	0.58
	Whole trading day	0.66	0.68	0.70	0.70	0.58
Model E	Half-on trading day	0.66	0.68	0.70	0.70	0.58
	Whole trading day	0.66	0.68	0.70	0.70	0.58
Model F	Half-on trading day	0.67	0.69	0.71	0.72	0.61
	Whole trading day	0.67	0.69	0.71	0.72	0.61

Table4. The experimental results in terms of CC

As can been seen from Table 4, it is clearly to find out that the CC of the baseline model achieves ranging from 0.53 to 0.67 with an average of 0.63. Again as it is reveal in Table 3 which is the experimental results in terms of RMSE, from model A, model B and model C, the statistics of CC reveals that the models incorporation of sentiment indicators will slightly outperform the baseline model but not significantly. And the incorporation of both sentiment indicators performs the best. Different from the results in Table 3, by introducing STEF, the slight improvement of the forecasting performance is archived by comparing model A and model D. Coinciding with the result from Table 3, time window of information does not influence the performance of models in general, except in model D with time lag p=2.

In sum, by the analysis of the experimental results above, some implications can be drawn. First, when comparing the results of the baseline model with model A, model B and model C, which incorporates sentiment indicators, it is satisfying that the indicators could basically reduce the error and improve the correlation coefficient. So, it is a reflection that the indicators brought in from extracting sentiment from Twitter could be useful in improving the model in predicting stock price. However, the slight improvement showed cannot significantly change the status of the model to make it more plausible and convinced. Yet it could safely draw a conclusion that the proposed method can quickly and easily capture human behavior, and improve the performance of financial market prediction. Second, further examinations of utilizing STEF to integrate the sentiment on the consequences of model D, model E and model F, though the experimental results outperform the baseline model, they are not obvious when compared with model A, model B and model C. So, maybe the general calculation is useful enough. As for the reason why the STEF performances are mediocre, it could be explained that maybe time window is so short a time that any big event happen at a day could equally make influences on the closing quotation. It may seem a little surprising but actually harmony with the reality. Finally, when time window is extended to the whole trading day, great improvement from the model A, model B, model C, model D, model E and model F can not be found. However, the discovery could be put into practical use for it inks us that the information made during the opening trading time could be necessary for predicting. It embraces the efficient market hypothesis that market itself could absorb the information before quickly and efficiently.

5 CONCLUSIONS

This paper proposes an online sentiment based forecasting method for financial market prediction using web mining. In terms of the experimental results, it is revealed that by adding online sentiment, the forecasting performance can be improved in financial market prediction. However, by introducing STEF, the forecasting performance does not have obvious improvement. It implies that maybe the general calculation is useful enough, or in some special situation, STEF can be in effect, which can be

further investigated in future work. As a whole, the empirical analysis indicates that the proposed online sentiment based forecasting approach can be used as a potential alternative to forecast financial market trend.

In addition, this study also remains some research questions for further consideration. First of all, this paper employs the simple sentiment classification (e.g. positive and negative) for financial market prediction, so some rigorous emotion theories can be offered to research the relationship between online sentiment and financial market trend. Secondly, the twitters are not connected to a specific stock but the S&P index, and the twitters are not filtered to finance related ones. Hence the study only reflects the relationship between general public emotion and the S&P index. The more detail experiments on specific stocks and their relative twitters are needed for further exploration. Thirdly, some special scenarios, such as voting and financial crisis, can be considered in financial market prediction, because in these situations, human behaviour may affect the market more heavily. Finally, in the process of extracting online sentiment, as the impact of each online user is different, the impact of online users can be considered for gathering the public sentiments.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable suggestions which have helped greatly in improving the quality of this paper. This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 71001103) and Beijing Natural Science Foundation (No. 9122013).

References

Bildirici, M. and Ersin, O.O. (2009). Improving forecasts of GARCH family models with the artificial neural networks: an application to the daily returns in Istanbul stock exchange. *Expert Systems with Applications*, 36, 7355–7362.

Bollen, J., Mao, H. and Zeng, X.J. (2010). Twitter mood predicts the stock market. *The Working Paper*, 1-7.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. *IEEE Computer*, 91-94.

Cao, Q., Leggio, K.B. and Schniederjans, M.J. (2005). A comparison between Fama and French's model and artificialneural networks in predicting the Chinese stock market. *Computers and Operations Research*, 32, 2499-2512.

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

Chiu, D.Y. and Chen, P.J. (2009). Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. *Expert Systems with Applications*, 36, 1240-1248.

Dacorogna, M.M., Gauvreau, C.L., Muller, U.A., Olsen, R.B. and Pictet, O.V. (1996). Changing time scale for short-term forecasting in financial markets. *Journal of Forecasting*, 15, 203-227.

Das, S.R. and Chen, M.Y. (2007). Yahoo! for Amazon: sentiment extraction from small talk on the web. *Management Science*, 53(9), 1375-1388.

Hsu, C.M. (2011). A hybrid procedure for stock price prediction by integrating self-organizing mapand genetic programming. *Expert Systems with Applications*, 38(11), 14026-14036.

Hung, J.C. (2009). A fuzzy GARCH model applied to stock market scenario using a genetic algorithm. *Expert Systems with Applications*, 36, 11710-11717.

Kanas, A. and Yannopoulos, A. (2001). Comparing linear and nonlinear forecasts for stock returns. *International Review of Economics and Finance*, 10, 383-398.

Lee, R.S.T. (2004). IJADE stock advisor: an intelligent agent based stock prediction system using hybrid RBF recurrent network.*IEEE Trans. System, Man, and Cybernetics - Part A, Systems and Humans*, 34(3), 421-428.

Li, X., Deng, X., Wang, F. ad Dong, K. (2010). Empirical analysis: news impact on stock prices based on news density. *IEEE International Conference on Data Mining*, 585-592.

Liao, Z. and Wang, J. (2010). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems with Applications*, 37, 834-841.

Lu, C.J., Lee, T.S., Chiu, C.C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2), 115-125.

Mitternayer, M.A. (2004). Forecasting intraday stock price trends with text mining techniques. *Proceedings of the Hawaii International Conference on System Sciences*, 1-10.

Nair, B.B., Mohandas, V.P. and Sakthivel, N.R. (2010). A decision tree - rough set hybrid systems for stock market trend prediction. *International Journal of Computer Applications*, 6(9), 1-6.

Olson, D. and Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, 19, 453-465.

Refense, A.P.N. and Holt, W.T. (2001). Forecasting volatility with neural regression: a contribution to model adequacy. *IEEE Trans. Neural Networks*, 12(4), 850-864.

Rodriguez-Gonzalez, A., Garcia-Crespo, A., Colomo-Palacios, R., Lglesias, F.G. and Gomez-Berbis, J.M. (2011). CAST: Using neural networks to improve trading systems based on technicalanalysis by means of the RSI financial indicator. *Expert Systems with Applications*, 38, 11489-11500.

Schumaker, R.P. and Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing and Management*, 45, 571–583.

Schumaker, R.P. and Chen, H. (2010). A discrete stock price prediction engine based on financial news. *IEEE Computer*, 51-56.

Sehgal, V. and Song, C. (2007). SOPS: stock prediction using web sentiment. *Seventh IEEE International Conference on Data Mining*, 21-26.

Tseng, F.M., Yu, H.C. and Tzeng, G.H. (2002). Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change*, 69, 71-87.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Wang, Y.F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications*, 22, 33-39.

Xu, W., Zuo, M., Zhang, M. and He, R. (2011). A neural network-based ensemble forecasting method for financial market prediction. *International Journal of Advanced Mechatronic Systems*, 3(4), 259-267.

Yu, L.,Chen, H.,Wang, S. and Lai, K.K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Trans. Evolutionary Computation*, 13(1), 87-102.