Web Intelligence 14 (2016) 287–300
DOI 10.3233/WEB-160345
IOS Press

A topic-based sentiment analysis model to predict stock market price movement using Weibo mood

Wenhao Chen [a,*], Yi Cai [b], Kinkeung Lai [c,d] and Haoran Xie [e]
[a] *Department of Management Science, City University of Hong Kong, Hong Kong*
E-mail: wenhachen2-c@my.cityu.edu.hk
[b] *School of Software Engineering, South China University of Technology, Guangzhou, China*
E-mail: ycai@scut.edu.cn
[c] *Department of Industrial and Manufacturing Systems Engineering, Hong Kong University, Hong Kong*
E-mail: mskklai@cityu.edu.hk
[d] *International Business School, Shaanxi Normal University, Xian, China*
[e] *Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*
E-mail: hrxie2@gmail.com

Abstract. Over the past several years, as the development of Internet, social media websites such as Twitter and Weibo have received much attention due to their enormous users. A lot of research has been done on sentiment analysis and opinion mining in these websites. However the number of research on using the data in the social media websites to predict the stock market price movement is limited. Behavioral economics and behavioral finance believe that public mood is correlated with economic indicators and financial decisions are significantly driven by emotions. This paper first presents a Chinese emotion mining approach and discusses whether the public emotions or opinions in the Chinese social media websites could be used to predict the stock market price in China. The experimental results demonstrate that the emotions automatically extracted from the large scale Weibo posts represent the real public opinions about some special topics of the stock market in China. Some public mood states extracted such as the "Happiness" and "Disgust" states are highly correlated with the change of stock price according to the Granger causality analysis. Finally, a nonlinear autoregressive model with exogenous sentiment inputs is proposed to predict the stock price movement.

Keywords: Sentiment analysis, public mood, stock price forecasting, artificial neural network

1. Introduction

For many years, a lot of research has been conducted on the stock price prediction [10]. Early research focuses on efficient market hypothesis and random walk theory [8] which indicates that the stock prices are driven by new information such as news rather than previous prices. Nowadays many different kinds of model have be developed to forecast the stock price [17,29,36]. And some recent research indicates that although news are unpredictable however according to the behavioral finance, the early indicators could be detected through different approaches. Many researchers suggest to use online social media (blogs, micro-blogs, facebook etc.) to obtain these indicators and improve the forecasting performance [3,24].

The number of social media website users around the world has a high rapid growth over the past 10 years. The exponentially increased user-generated content in social media websites such as Facebook and Twitter has provided a potential data source for opinion mining as users share their opinions in these websites by text posts, pictures and even videos. In China, Weibo is one of the largest microblog social media websites and the platform is similar as Twitter. The

*Corresponding author. E-mail: wenhachen2-c@my.cityu.edu.hk.

distinctive nature of social media websites makes it a valuable source for mining public opinions or emotions. Firstly, a lot of users post their opinions about different topics in the website which represent their emotions in some extent. Secondly, these posts are highly time sensitive. Active users in these websites post their opinions every day, even every hour. Thirdly, some social media websites such as Weibo have an approval mechanism for registration which makes sure each user in these systems is a real person. Fourthly, some organizations and companies use social media websites such as Facebook, Twitter and Weibo as their formal channel to publish their opinions to the public. As a result the data collected from these social media websites can help researchers to understand the public mood states or emotions.

Recently significant progress has been achieved in extracting states of public mood directly from social media websites such Twitter and use them to predict the stock price movement [14]. However, the number of research on Chinese social media websites is limited and the number of research on how to use the textual sentiment analysis result of social media websites towards Chinese stock market is even lower. In recent years, China has already been the second largest economies and the stock market in China has a great impact to the global stock markets. The intraday trading volume on Shanghai Stock Exchange (SSE) has been more than 1 thousand billion RMB on 20th April 2015 which is more than any stock exchange intraday trading volume in the history. Instead of Shanghai Stock Exchange, there are also Shenzhen Stock Exchange and Hong Kong Stock Exchange in China. How the stock market price in China be affected by the public mood change is important for economists, traders and socialist.

This paper is interested in discussing whether the public mood states extracted from Chinese social media websites such as Weibo can represent the real opinion of publics towards the stock market and how the public mood states impact the stock market. Although researches on behavioral economics and finance have already demonstrate the correlation between emotions and the stock price, this paper will discuss whether Weibo is a valid source to monitor the public mood states about the stock market and demonstrate the causality between the Weibo public mood states and the stock price. This paper first presents a topic-based approach to extract public mood states from Weibo. After that a Granger causality analysis is used to investigate the hypothesis that the extracted public mood states are predictive of the stock price movement in China. The experimental results indicate that some public mood states such as the "Happiness" and "Disgust" states are highly correlated with the movement of stock price. Finally, the nonlinear correlation between the stock price movement and the public mood in Weibo is discussed. A topic-based nonlinear autoregressive model with exogenous Weibo mood input is then proposed to predict the future stock price movement. The prediction accuracy of the price movement direction is competitive and high. The model could be used as a reference for the investment decision making process.

2. Background

In recent years, social media websites, such as Facebook, Twitter and Weibo become more and more popular. The user number is enormous and there are different user communities with diversified natures in these websites. A lot of data is published by the online users including comments and news in these websites. The data, to some extent, represents the public emotions or opinions. An increasing number of research has attempted to integrate sentiment analysis result of the online data into different models.

2.1. Public mood in social media websites

Social media is known as a computer-mediated tool which allows people to create, share or exchange information. It has many different forms including blogs, microblogs, photo sharing, video sharing, forum, tagging system and social network. The increasing popularity of social media websites and Web2.0 has led to exponential growth of user-generated content, especially text content on Internet. Abbasi et al. [1] have demonstrated that the information retrieval and automated analysis technique are useful for understanding the online content such as forum posts and social interactions in online communities. Through analyzing the blog content, Liang et al. [22] indicated that a company can get the first hand knowledge or feedback from its clients. And it can also help to understand how the online customer networks appear and evolve [5]. This kind of information extracted from blogs enable organizations to make better decision on critical business area which is important for business intelligence [27]. In terms of integration of social media sentiment to applications in different industries, a lot of research is

conducted to use the sentiment from twitter to forecast spikes in book sales [14] and the revenues of box-office for movies in North America [24]. Another research area for using text sentiment analysis result is the recommendation system. Many experiments have been done to integrate sentiment analysis results of online text into recommendation systems [32].

2.2. Behavioral economics and finance

Behavioral economics and finance study the impact of psychological or emotion factors on the related decision making area and the effect on stock price, returns and the risk of the market. Psychological research has already proved that emotions, in addition to information has a significant effect in human decision-making [7]. Behavioral finance has further demonstrated that investment decision of investors is more likely driven by their emotions [26]. And the momentum generated from the public mood with other factors such as economic factors determine prices of the stock market.

Early research on stock market prediction focus on building models with random walk theory and Efficient Market Hypothesis [9]. However many researches show that the stock market prices do not follow a random walk and some researchers suggest that some indicators of the price can be extracted from the online social media. Schumaker & Chen [31] applied machine learning methods to financial news articles and find that the sentiment in news articles has an immediately impact on the market price. The prediction model has the best performance by adding the news article factors. Based on the community sentiment retrieved from the posts on the Yahoo Finance Forum through an expert classification system, Liu et al. [23] has indicated the correlation between the sentiment and the stock price. Gibert [13] using the LiveJournal as a source, has extracted the anxiety, worry and fear mood from the posts. He found out that the increase on expression of anxiety have indicated that the S&P 500 Index will move downward soon. Bollen et al. [3] has investigated the correlation between the collective mood states from large-scale twitter feeds and the value of the Dow Jones Industrial Average (DJIA) over time. Using twitter posts as well, Zhang et al. [35] found out that the emotions can be used to predict NASDAQ and S&P500 index as well. Li et al. [20,21] discussed the problem in using News or summarization of News to predict stock price in Hong Kong. Rao et al. [30] propose a topic-level maximum entropy (TME) model for social emotion classification over short text.

2.3. Chinese social media analysis

As Chinese is quite different with English and it is more complicated in terms of recognition, segmentation and analysis, the number of research on Chinese social media mining is limited and the number of research on using the textual sentiment analysis result to predict the stock market in China is even lower. In previous research, Gao et al. [12] indicated the difference between Sina Weibo and Twitter. Yang et al. [34] have proposed a classifier to automatically detect the rumors from a mixed set of information from the posts in Sina Weibo. For sentiment analysis, Rui et al. found out that the correlation of anger among users is significantly higher than that of joy in Sina Weibo [11]. There are also researches about multi-lingual sentiment analysis, K. Ahmad et al. [2] developed a local grammar approach which works equally on English, Chinese (Sino-Asiatic) and Arabic (Semitic).

As the stock market in China including the Shanghai Stock Exchange, Shenzhen Stock Exchange and Hong Kong Stock Exchange has already been one of the largest stock market around the world. The turnover or trading volume in these stock exchanges is enormous and the volatility of the stock price is high. As a result, the risk of trading in these exchanges is quite high especially for the Shanghai Stock Exchange and Shenzhen Stock Exchange as they are still be considered as the emerging markets in which the regulation and rule is not mature and the stock price is more easily controlled by inside information, special events and public emotions. Accordingly, it is an important research topic that to find out how the price movement is affected by different behavioral factors especially the public emotions. This is the area that this paper want to discuss and the paper demonstrates that the public mood derived from Weibo is correlated with the stock price movement in China. In addition, at the end of this paper, a Chinese sentiment analysis approach to extract public mood states from Weibo and a nonlinear autoregressive exogenous model using Weibo mood to predict the stock price movement are proposed.

3. Data and system framework overview

Some research has already been conducted on using the overall public mood from Twitter to predict the stock market price [3]. However a lot of unrelated posts are in these social media websites. If researchers want to discuss the emotion impact on a stock, they need to find out the text which is highly related to the

stock. For example, even most users have bad comments on a movie in Twitter that should not have a big impact on the stock price of an oil company. As a result, to filter the unnecessary noise, this paper uses a topic-based approach to analyze the public mood, and discuss how users' comments or emotions about these topics affect the stock market. Because of the characteristic of social media websites, discussions and news in these websites such as Twitter and Weibo are topic based. For example, a lot of posts are published to discuss the impact of "Shenzhen-Hong Kong Stock Connect (SZ-HK connect)" in Weibo. In Twitter, users use hashtags to identify different topics for discussion. Extracting the public mood states or opinions based on topics has two advantages. First, it can filter out unrelated information and only focus on opinions about the specific topic. Second, the public mood states about some special topics are easily be reused to discuss the relationship between different topics and data. One of the data source discussed in this paper is the text content from a Chinese social media website called Weibo. It was launched in August 2009. As of mid 2014, there are 167 million monthly active users, more than 25 million posts each day.

The database used in this paper is a collection of public Weibo posts that was recorded from Jan 2015 to Feb 2016. For each post the database provides an identifier, the datetime of the submission and the text content. All these posts are crawled for Weibo.com. After that the posts are classified by date and stopwords and punctuation are removed for each post. As this paper is discussing the public mood related to Chinese stock market, 3 topics for discussion are selected which are: "Shanghai-Hong Kong Stock Connect (SH-HK connect) and Shenzhen-Hong Kong Stock Connect (SZ-HK connect)", "Interest rate and reserve rate cut in China", "Greece Government-Debt Crisis". The reason why these 3 topics are selected is that from 1st Jan 2015 to 31th Dec 2015, they are mostly discussed topics related to the stock market in the website. SH-HK connect has been launched on November 2014 and SZ-HK connect is planned to launched in 2016. It allows the individual and institutional traders in Shanghai or Shenzhen to trade the stock market in Hong Kong through the Connect. And the traders in Hong Kong also can use it to trade the stock market in Shanghai or Shenzhen. It provides different choices and an arbitrage chance for the traders in Hong Kong and Mainland China. It is considered as a great milestone in the development of China stock market and people believe that it will have a positive impact on China stock market. Greece Government-Debt Crisis is the debt crisis that happens in European Union (EU) in 2015, people worry that if Greece cannot pay the debt on time, it will cause the financial crisis in EU and the international crisis will be triggered again. Interest rate and reserve rate cut means the central bank in China plan to cut the bank lending interest rate and the amount of cash banks must keep in reserve. It will increase the flow of money in the economy and reduce the cost of financing to support the investment and developments of the real economy. As it will help to boost the growth of economy, it has a positive impact on the stock market. The second source of data is the daily end price of stocks in Hong Kong and Shanghai stock exchanges which could be downloaded from public available websites such as Yahoo!Finance and http://finance.sina.com.cn/.

The system framework can be summarized as follows. In the first phase, it will filter the posts and only extract the posts containing the keywords related to the 3 topics. Posts are grouped under the same topic. In the second phase, it will use two tools to measure the sentiment score for different emotions in the posts. The first tool, Jieba, will help to analyze the Chinese text and segment the text into meaningful Chinese words. It is an useful tool to segment user-generated content in forums [33]. The second tool, Chinese Emotion Word Ontology (CEWO), is a Chinese sentiment word dictionary which classifies the Chinese emotion words into 7 different categories [6]. A novel method to generate the sentiment score or mood states for each post based on the CEWO is applied to the dataset. After that it combines the sentiment score for all posts on the same day under the same topic. As a result, the daily sentiment score for different emotions are generated for each topic. The sentiment score across the observation time can be regarded as time series. In the third phase, it investigates the hypothesis that public mood measured by our mechanism is predictive of future stock price using Granger causality analysis. As the public mood generated is related with different stocks, the impact on different kinds of stocks are discussed. At the end, a nonlinear autoregressive model with exogenous Weibo mood inputs is proposed to predict the price and moving directions of a stock.

4. Data processing: A new Chinese sentiment analysis method

As mentioned above, the raw text data is restored in database first. And then the Chinese segmentation tool

English	Chinese
Happiness	乐
Good	好
Anger	怒
Sadness	哀
Fear	惧
Disgust	恶
Surprise	惊

Fig. 1. The 7 Categories in the Chinese Emotion Word Ontology.

Jieba is used to segment the text into different words. The Jieba tool use the dynamic programming to find out the most probable combination based on the word frequency. The segmentation result will be restored in the database as well. The next step is to generate the sentiment score or mood states from each post. For this part, the model will use the Chinese Emotion Word Ontology constructed by the IR lab in Dalian University of Technology. This ontology is constructed based on Ekman theory of 6 basic emotions (Happiness, Sadness, Surprise, Fear, Disgust and Anger). And one more emotion "good" is added to make it more comfortable for Chinese language analysis [6]. Figure 1 shows the mapping from English to Chinese for the 7 categories of emotion words. These categories are also known as different dimensions of public mood. Each emotion word in the Ontology has its own category tag such as "Happiness", intensity value and polarity value.

The proposed method to generate the sentiment score or public mood states including 4 steps. First, Jieba is used to segment the post and translate the post to a list of meaningful words including different lexical class such noun, verb, adv and adj. Secondly, it will process the word list and search each word of the list in the Chinese Emotion Word Ontology based on its lexical class. If a word i is found in the Ontology, the intensity value of the word in the Ontology will be recorded as s_i. Thirdly, the sentiment score of the post will be calculated as follows:

$$y = \max_{j=1,\ldots,7}\left(\sum_{i\in I(j)} s_i\right) \tag{1}$$

Where $I(j)$ is a set including all words found in jth emotion category of the Ontology for a post, $I(1)$ means the words found in the first emotion category of the Ontology which is the category of "Happiness". The algorithm assumes that each post only presents one kind of emotion which is the one that has the overall highest intensity value or sentiment score. As a result each post in the database will be attached an tag which identifies the category it belongs to such as "Happiness". In the fourth step, the frequency of the posts belong to the same emotion category is calculated for each day. The value will be represented as $x_{i,t}$ where i belongs to one of the 7 emotion categories and t means date t in the observation period. Then the time series for emotion i can be represented as X_i.

After processing the raw text data of Weibo posts, for each topic discussed, the model generates 7 time series. Each time series represent the movement of one dimension of the public mood or emotion such as "Happiness". Another data source is the time series of the stock price. The time series of the stock close price is downloaded from Yahoo!Finance and other public websites. As the emotion time series are associated with different stocks. the close prices from 1st Jan 2015 to 31th Mar 2016 for 4 different stocks including the Hang Seng Index (HSI), Shanghai Stock Exchange Composite Index (SSECI), Hong Kong Exchanges and Clearing Limited corporation stock (stock code is 388) and Shanghai Connect Index (stock code is 000159) are downloaded.

5. Public mood validation

To validate the ability of the proposed method to capture various aspects of public moods, the algorithm is applied to the Weibo posts published from 1st Jan 2015 to 31th Mar 2016. The distribution of the posts indicates that the main part of the positive topic posts such as "interest rate cute" are the posts with the positive emotion such as "Happiness" and "Good" and vice versa. The distribution of the posts is shown in Fig. 2.

As mentioned, the interest rate and reserve rate cut will booth the economy and the stock market. Actually, the China government has cut the interest rate or reserve rate three times during 1st Jan to 31 July 2015 (on 28th Feb, 10th May, 27th June). As a result, most posts about the interest rate and reserve rate cut locate in the positive emotions such as "Happiness" and "Good" which is 85% of all related posts. The distribution of the posts related to SH-HK and SZ-HK connect is similar. However the distribution of the posts related to Greece government-debt crisis is different. As a lot of people is worried about the future of the European Union financial system and whether the Greece crisis will cause an international government debt crisis, there is almost 27% posts is classified as "Disgust" emotion. There are

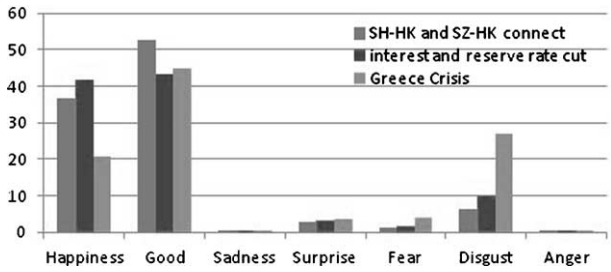

Fig. 2. The percentage distribution of the posts with different emotions under 3 topics.

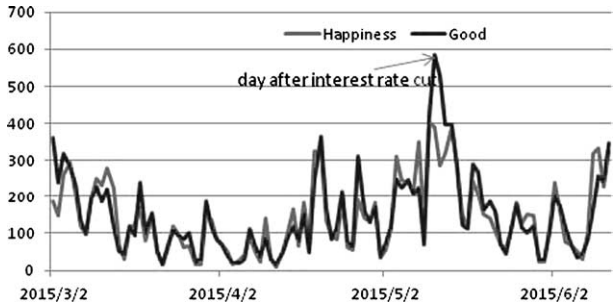

Fig. 3. Public mood states between Mar to June 2015 shows public responses to the announcement of interest rate cut.

also a lot of posts locating in the "Good" category as the Greece government-debt problem is solved in July. As the posts are mainly distributed in the three categories:"Happiness", "Good", "Disgust", this paper will discuss the correlation between the stock market and these three emotions.

Figure 3 shows the two public mood time series after the data processing which are "Happiness" time series and "Good" time series for the topic "interest rate and reserve rate cut". On 10th May, China central bank announced a new round of interest rate and reserve rate cut which is an event that may have a unique, significant and complex effect on the public mood. The result demonstrates that the time series generated through the approach successfully identifies the public emotional response to the announcement of interest rate cut. It has a significant but short-lived up tick in positive sentiment "Happiness" and "Good" on that day. Actually after the announcement, the sentiment score increases a lot sharply. In addition, before the announcement the sentiment score starts to climb up as there are some rumors appearing in the market which mentioned that the government will cut the interest rate soon and the expectation of the new round interest rate cut becomes higher.

To visualize the correlation between the public mood time series and the stock price movement, the data set for the topic "Shanghai-Hong Kong Stock Connect (SH-HK connect) and Shenzhen-Hong Kong Stock Connect (SZ-HK connect)" for the period 1st April to 31th May is used. The reason for using this time period is that on 8th April, Chinese investors, for the first time, used the entire 10.5 billion RMB daily quota in a cross-border programme buying Hong Kong stocks, boosting turnover under the Shanghai-Hong Kong Stock Connect to a record. The buying makes the Hang Seng Index increase 3.8 percent to its highest level in nearly seven years.

Two time series are plotted in Fig. 4. First one is the daily close price of Hang Seng Index. And the Second one is the time series for the public mood state "Happiness". To normalize and compare the value in these two time series, the z-score of their delta value is calculated. As shown in Fig. 4, both time series frequently overlap or on the same direction. The movement of the two time series is similar according to their peaks and bottoms. On 8th April, both time series climb to the peak with almost the same z-score which is more than 1.5 standard deviations comparing to the average (Hang Seng Index: 1.74 and Happy Mood: 1.68). These behaviors of the time series demonstrate that the "Happiness" time series generated by the approach correctly indicate the public mood change to some positive events and is related with the Hang Seng Index price. The correlation between these two time series will be discussed in more details in next section.

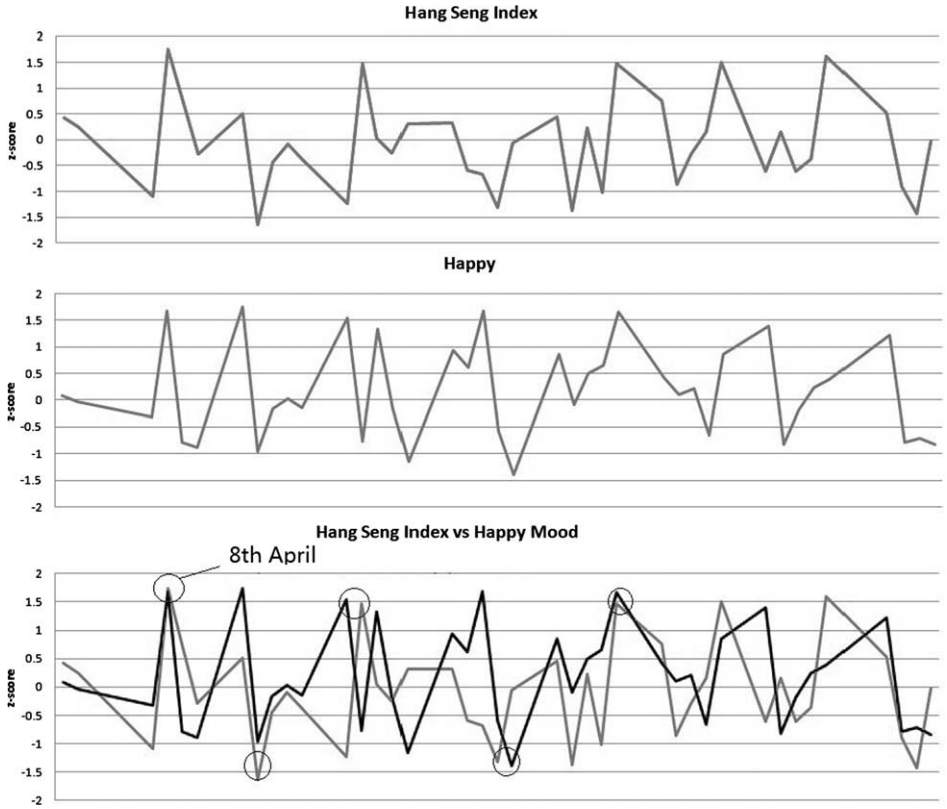

Fig. 4. The z-score of the Hang Seng Index delta price vs the z-score of the Happiness time series.

6. Granger causality analysis

Many economic time series are unit root process [25]. Therefore, the Augmented Dickey-Fuller (ADF) test is applied to each time series for identifying the unit root process. As the generated time series are proved to be unit root process, first differencing method is used to minimize the chance of spurious relationships. The public mood time series that is not covariance stationary is converted to D_i,

$$D_{i,t} = \Delta x_{i,t} = x_{i,t} - x_{i,t-1} \tag{2}$$

To avoid the effect of various magnitudes, the public mood time series D_i is normalized by computing the z-scores:

$$Z_{i,t} = \frac{D_{i,t} - \bar{D}_i}{\sigma_i^2} \tag{3}$$

Where $\bar{D}_i$ is the average of D_i and σ_i is the standard deviation of D_i over time period n. In the experiment, as it will discuss the moving average of stock price, n is defined as 10 and $\bar{D}_i$ is the 10 days moving average of the sentiment score change.

The time series of stock prices is defined to reflect the daily changes in stock market value, i.e, the delta value between the close price of a stock in day t and day $t-1$. Z-score computing is also applied to the delta value which helps to standardize the stock price change. $P_{i,t}$ equals to the z-score value of the delta. After establishing the public mood time series, to discuss whether variations of the public mood states affect the price in the stock market, this paper will apply the Granger causality analysis to the time series generated by the new Chinese sentiment analysis method, Z_i, and the stock close price time series, P_i. This paper performs the Granger causality analysis according to the linear model as shown in Eq. (4) for the period of time between 1st Jan to 31st July 2015. As there are three topics to discuss and these topics are related to different stocks, different linear regression experiments between specific stock price time series and the public mood time series will be conducted.

Table 1
The linear regression analysis result between the "Happiness" time series and the close price movement of Shanghai Connect Index

Lag	Std Error	t stat	P-value
1 Day	0.096308	−0.646502	0.519301
2 Day	0.098495	−1.300428	0.196172
3 Day	0.101275	−2.219358	0.028515
4 Day	0.103179	−0.251972	0.801533
5 Day	0.101925	−0.225403	0.822084
6 Day	0.098601	0.961458	0.338431
7 Day	0.093337	1.237520	0.218529

$$P_t = \alpha + \sum_{i=1}^{n} \beta_i P_{t-i} + \sum_{i=1}^{n} \gamma_i Z_{t-i} + \varepsilon_t \qquad (4)$$

Where P_t is the z-score of the stock day close price difference between day t and $t-1$, Z_t is the z-score of the sentiment score change of a specific public mood dimension or emotion such as "Happiness". The value of n is set to 7, as previous research [3] has demonstrated that public mood on day $t-3$ has the highest correlation with stock price on day t.

First the topic "SH-HK and SZ-HK connect" is discussed. Table 1 presents the linear regression result between the time series for the close price of Shanghai Connect Index (stock code: 000159) and the time series of the emotion "Happiness" related to the topic "SH-HK and SZ-HK connect". Shanghai Connect Index is the composite index including all stocks that can be traded through the SH-HK connect by international traders in Hong Kong. As a result, the price movement of the Shanghai Connect Index is directly affected by traders' opinions or emotions about SH-HK connect. The result shows that the lag value of Z_{t-3} has a P-value equal to 0.028515 which is lower than the significance level 0.05. Based on the result of the Granger causality, the null hypothesis that the mood series do not predict the price movement of Shanghai Connect Index can be rejected, i.e. $\gamma_{1,2,3,\ldots,7} \neq 0$. However only the lag with 3 days has significant causal relations with price movement of the Index which is similar as Bollen's research [3] on Twitter post. In his research, he mentioned changes of public mood match shifts in the stock price that occur 3–4 day later. This behavior can be explain as that it may take some time for the public mood in social media websites to gain a momentum that will affect investors decision in the real market. Other 6 emotions time series are not significantly correlated with the time series of Shanghai Connect Index according to the p-value.

Table 2
The p-value of linear regress analysis result between "Happiness" time series and 3 different stocks

Lag	000159	388	HSI
1 Day	0.519301	0.096123	0.538607
2 Day	0.196172	0.546172	0.428106
3 Day	0.028515	0.458417	0.045231
4 Day	0.801533	0.010188	0.155256
5 Day	0.822084	0.130202	0.198888
6 Day	0.338431	0.220351	0.500008
7 Day	0.218529	0.934022	0.301673

To further demonstrate the forecast ability of the "Happiness" time series to other stocks, the p-value of the Granger causality correlation with 2 more stocks are calculated as well. The 2 stocks are HSI and the Hong Kong Exchanges and Clearing Limited corporation stock (stock code is 388). They are highly related to the topic "SH-HK and SZ-HK connect". HSI is the composite Index of Hong Kong stocks which will be affected by the SH-HK connect as more capitals can be invested on Hong Kong market. Hong Kong Exchanges and Clearing Limited Company is the company that provides the service of SH-HK connect. More trader are investigating through the SH-HK connect, more commissions it can receive. As it can gain more profits, the valuation of the stock is changed and the price will be increased as more investors will be attracted. The results are shown in Table 2. According to the p-value 0.045231, the lag value of Z_{t-3} has significant correlation with the price movement of HSI. Similar as HSI, the price movement of 388 is significantly correlated with the lag value of Z_{t-4}. The p-value is 0.010188 which is much lower than the significance level 0.05. The null hypothesis can be rejected and the "Happiness" time series is predictive of the stock price movement including Shanghai Connect Index, HSI and the stock of Hong Kong Exchanges and Clearing Limited.

To demonstrate the predictive ability of public mood states with different topics, the Granger causality analysis is also applied to other topics. For the topic "Greece Government-Debt Crisis", the public mood time series other than "Disgust" is not significant correlated with the price movement of HSI. Only the lag value Z_{t-2} of the time series of "Disgust" has significant causal relation with the price movement of HSI considering the significance level 0.1. The experiment result is shown in Table 3.

For the topic "interest rate and reserve rate cut", as it is a method that China government used to booth the

Table 3
The linear regress analysis result between "Disgust" time series about"Greece Government-Debt Crisis" and the close price of HSI

Lag	Std Error	t stat	P-value
1 Day	0.131315	−0.60719	0.546991
2 Day	0.146389	−1.86881	0.068634
3 Day	0.1709	0.635789	0.528363
4 Day	0.176888	1.476039	0.147392
5 Day	0.17898	1.233007	0.22443
6 Day	0.166583	0.024174	0.980829
7 Day	0.146273	−0.21275	0.832549

Table 4
The linear regress analysis result between public mood states related to "interest rate cut" and the close price of SSECI

Lag	Happiness	Good	Disgust
1 Day	0.254616	0.515584	0.395321
2 Day	0.111199	0.455075	0.041898
3 Day	0.007864	0.0676	0.008356
4 Day	0.944177	0.201313	0.959366
5 Day	0.858916	0.925238	0.995641
6 Day	0.8697	0.555839	0.193032
7 Day	0.355159	0.069943	0.384485

China stock market, the paper will discuss the correlation between the public mood time series and the close price of Shanghai Stock Exchange Composite Index (SSECI). The results are shown in Table 4. According to the p-value (0.007864), the lag value of Z_{t-3} for emotion "Happiness" is highly correlated with the price movement of SSECI considering the significance level 0.05. And for the "Good" time series, considering the significance level 0.1 and p-value 0.0676, the lag value of Z_{t-3} is also correlated. In addition, the lag value of $t-2$ and $t-3$ of the "Disgust" time series have significant causal relationship with the SSECI value as well as the p-value is 0.041898 and 0.008356. The experiments are not testing the actual causation but whether one time series has predictive information about the price movement or not. The result shows that the null hypothesis that the stock price movement is not correlated with the change of public mood extracted from Weibo can be rejected with high level confidence and the time series of "Happiness", "Good" and "Disgust" could be used to predict the price movement of SSECI. The reason why only "Happiness", "Good", "Disgust" are correlated with the stock price is because of the characteristics of Weibo. Most of the post are related and classified to these 3 dimensions through the proposed new Chinese sentiment analysis model. Another reason is that investors' decision is more likely be affected by strong emotions which has high polarity such as "Happiness" and "Disgust". In addition, the spread of these strong emotions such as "Happiness" in the social media websites is more quickly as it will attract more attentions than other kinds of emotions. As a result, there will be more posts with the same strong emotions.

7. A nonlinear autoregressive model with exogenous sentiment inputs for stock prediction

Although Granger causality analysis result has proved that certain public mood dimensions are correlated with the specific stock price movement in China, the analysis is based on linear regression. However the relation between public mood and the stock market is more likely to be non-linear. The ability of artificial neural network in solving the non-linear time series modeling issue has already been proved [16,18]. Some researchers have already applied the neural network to the topic of stock price prediction [17]. Zhu et al. indicate that neural network models augmented with trading volumes leads to improvements in forecasting performance under different terms of forecasting horizon [36]. In addition, some research has been done to forecast the exchange rate using some specific type of neural network such as self-organizing dynamic neural network [19]. In previous literatures [36] trading volume is considered as an augmentation to improve the performance of neural networks. In this paper, whether the public mood states from Weibo could reinforce the forecasting ability of neural network will be discussed.

7.1. Model set up

The overview of the nonlinear autoregressive model with exogenous sentiment inputs is shown in Fig. 5. Two data sets are required in the model. one is the historical stock price which can be downloaded from the Internet, another one is the public posts set from Sina Weibo. As the purpose of this model is to understand the moving direction of the stock price and predict the future trend, the raw price of a stock will be first transformed to the 5 days moving average of the stock price. Moving average is widely used in previous research to predict the trend of the stock price movement [4,15,28]. As a result, the first input time series is generated based on 5 days moving average, for example, $\{159_1, 159_2, \ldots, 159_t\}$ where t is the total number of days for the testing period.

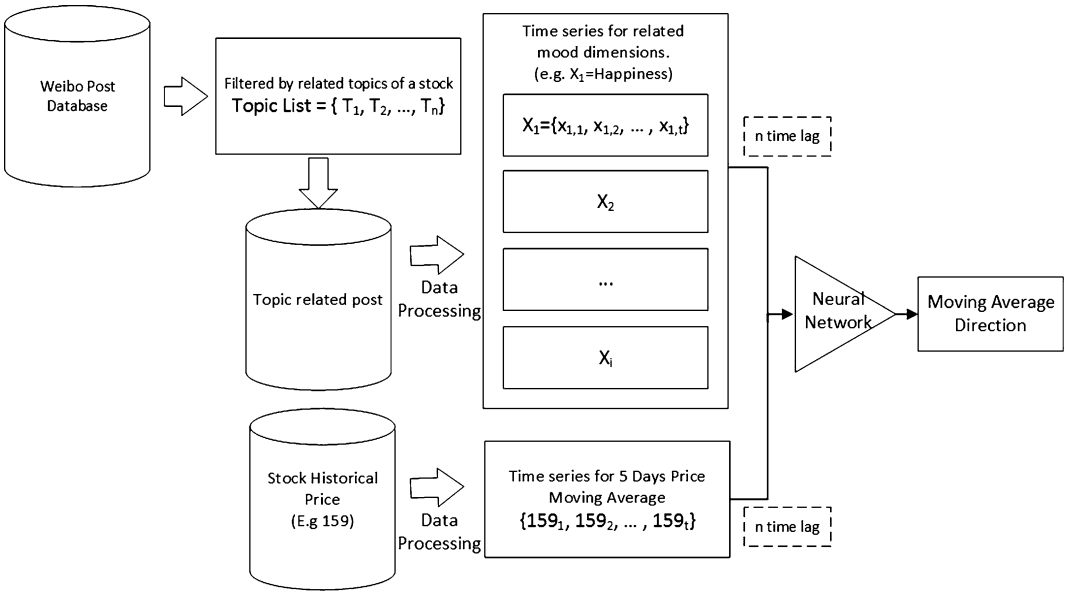

Fig. 5. The nonlinear autoregressive model with exogenous sentiment inputs.

Another part of the model is the topic based sentiment score time series generation. Through the Granger causality analysis result in last section, the stock price is proved to be related with the mood indicators for some specific topic. According to the efficient market hypothesis theory, stock price is driven by new information. The new information in the proposed model here is considered as topic related. The new information for the specific topics which are related to a stock and the reactions of the public drive the price movement. To generate the sentiment score inputs, the first step is to find out all related topics about a stock. For example, the topic "Shanghai Hong Kong Connect and Shenzhen Hong Kong connect" is highly related to the Shanghai Stock Connect Index (000159). In addition, for each topic, the correlated keyword list $T_1 = \{k_{1,1}, k_{1,2}, \ldots k_{1,m}\}$ will be extracted from Sina Weibo. The model then use these keywords to filter the unnecessary post from Sina Weibo. After that, each topic will have an associated data set of posts to generate the topic based sentiment score. Finally the new Chinese sentiment analysis method presented in Section 4 is applied to calculate the daily sentiment score for each topic and generate the sentiment score time series for different mood dimensions.

The sentiment score will be normalized to the z-score as what have been done in Section 6. Through Granger causality analysis, unrelated mood dimensions will be filtered out, and only useful time series will be remained. For example for Shanghai Stock Connect Index (000159), through the above Granger analysis, "Happiness" time series about "Shanghai Hong Kong stock connect and Shenzhen Hong Kong stock connect" are proved to be related with the stock price. Accordingly it could be considered as an exogenous sentiment input in the proposed model. Other mood dimensions such as "disgust" will be ignored. Through the proposed Chinese sentiment analysis method in Section 4, for each topic, 7 time series are generated based on the 7 mood dimensions. However, not every dimension could be used. It is possible that all 7 dimensions are proved to be not related. So as shown in Fig. 5, the number of sentiment inputs, i, is not equal to the number of topics, n, associated with the stock. After getting all the related sentiment score time series, the proposed model will consider them together with the moving average time series as the model input and use a neural network to predict the future price.

A back propagation feed forward neural network with multi hidden layers is implemented to build up the nonlinear autoregressive exogenous model. To build up a unbiased comparison of model performance, the same parameter value of the neural network is maintained. To predict the moving average price of a stock on day t the historical price of the stock in the past n days and the past value of specific public mood time series are considered as input attributes. As the Granger analysis result in last section suggests that the most related time lag value is $t - 3$ or $t - 4$. In

Table 5
Stock price prediction result

Dataset	000159		388		Hang Seng Index	
Models	I_1	I_0	I_1	I_0	I_1	I_0
MSE	286.2919	460.3728	0.5796	0.9306	3646	6755.2
MAPE	0.0039	0.0064	0.0034	0.0044	0.0024	0.0033
Direction (%)	89.66	75.86	77.14	62.86	87.1	70.9

the neural network model n is defined as 5. For each hidden layer, the number of neuron is defined as 10 in the model. For training, Levenberg-Marquardt algorithm is adopted as it is considered as a generic method to solve non-linear least squares problems. Mean square error (MSE) is used to monitor and compare the performance of the model. The model has one input layer, one output layer and multiple hidden layers. As shown in Fig. 5, the inputs of the neural network are the moving average price time series and the different topic-based sentiment score time series, e.g. $\{159_{t-1}, \ldots, 159_{t-n}, x_{1,t-1}, \ldots, x_{1,t-n}, \ldots, x_{i,t-1}, \ldots, x_{i,t-n}\}$, and the output is the moving average price in the next step and the direction of the price movement (up or down). This model is topic-based and could be easily extended by adding new topics or sentiment inputs. As every topic has a life cycle, there is no fixed number of topics in the model. The topics list related to a stock is changed everyday. The model is self-adaptive. The topic which is not related to the stock could be removed. Or if a new information or news is announced, there could be another topic that can be included. This paper will not discuss how the topics list could be built. It is only interested in whether the model with exogenous sentiment inputs is useful or not.

7.2. Experimental results

For testing of the nonlinear autoregressive model with exogenous sentiment inputs, the training data set used is the collection of Weibo posts from 1st Jan 2015 to 31 Dec 2015. And the testing period is from Jan 2016 to March 2016. As discussed in Section 6, according to the Granger causality analysis result, the "Happiness" mood dimension time series derived from the topic "Shanghai Hong Kong stock connect and Shenzhen Hong Kong stock connect" are highly related to the stock price of 000159, 388 and HSI. Other time series are not related. As a result, in this experiment, only the "Happiness" time series is involved as an exogenous sentiment input for predicting the price movement of 000159,388 and HSI. Table 5 is the testing result of the model. To demonstrate the ability of sentiment inputs, two nonlinear autoregressive models are applied to the same data set, one has the "Happiness" sentiment score time series as an input, another doesn't. Through the result, it is clear that by adding sentiment time series as an input, the performance is better. The nonlinear models used in the experiment are shown in Eq. (5). I_0 indicates the normal autoregressive model. I_1 represents the proposed model with exogenous sentiment inputs. In the experiment, only the "Happiness" time series associated with the topic "Shanghai Hong Kong stock connect and Shenzhen Hong Kong stock connect", $X_{1,t-5,\ldots,t-1}$ is consider as the sentiment input. If more related topics and the corresponding sentiment time series could be detected, the performance may be better. However in this experiment, the target is to discuss the ability of the new model and whether the prediction result could by improved by adding the sentiment score time series, finding out all the related topics for a stock is not the concern of this paper.

$$\begin{aligned} I_0 &= \{159_{t-5,\ldots,t-1}\} \\ I_1 &= \{159_{t-5,\ldots,t-1}, X_{1,t-5,\ldots,t-1}\} \\ I_0 &= \{388_{t-5,\ldots,t-1}\} \\ I_1 &= \{388_{t-5,\ldots,t-1}, X_{1,t-5,\ldots,t-1}\} \\ I_0 &= \{HSI_{t-5,\ldots,t-1}\} \\ I_1 &= \{HSI_{t-5,\ldots,t-1}, X_{1,t-5,\ldots,t-1}\} \end{aligned} \tag{5}$$

As shown in Table 5, for the Shanghai stock connect index (000159), the mean square error for the testing period is 460.3728 using the general nonlinear autoregressive model as I_0. Comparing to that, the mean square error is decreased to 286.1919 by adding the "Happiness" time series as another input. Similar as 000159, the mean square error for the stock 388 and Hang Seng Index is decreased as well, from 0.9306 to 0.5796 for 388 and from 6755.2 to 3646 for HSI. According to the result, it can be concluded that by adding one sentiment time series as an input, the prediction performance is improved. For investment deci-

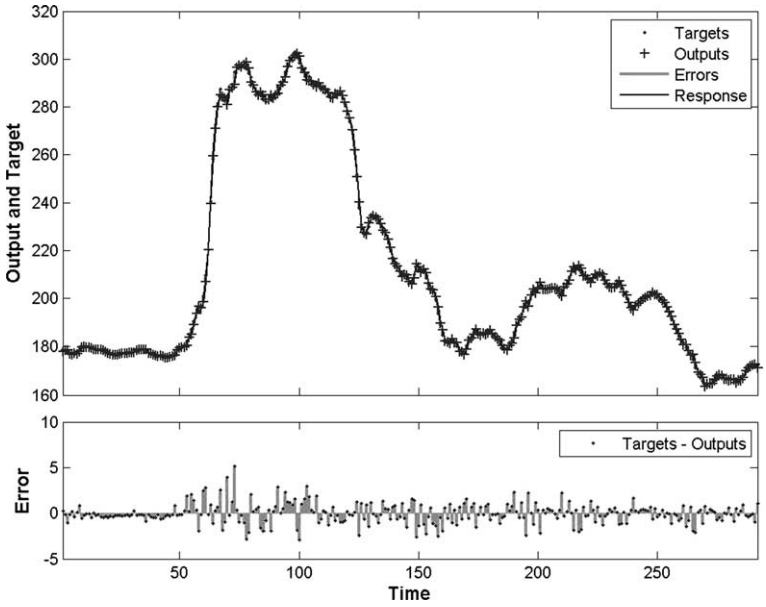

Fig. 6. Model outputs comparing to the expect values for 388.

sion making, it is important to know the future direction of a stock. Moving average represents the moving direction of a stock in a certain period. To demonstrate the ability of predicting the price movement direction for the proposed model, the directions determined by the model (daily price up or down) are compared to the exact directions representing in the history data of a stock. The accuracy of direction prediction is high according to Table 5. The highest accuracy is for 000159 which is 89.66 percent. The directions are not matched only for a few days. Same as the mean square error, the accuracy is increased by adding the sentiment time series as inputs, from 75.86 percent to 89.66 percent for 000159, from 62.86 to 77.14 percent for 388 and from 70.9 to 87.1 for Hang Seng Index.

Looking into the details of the prediction results, Table 6 has shown the precision, recall and F1 value for predicting up directions based on the proposed model. After adding the sentiment time series into the inputs, the precision, recall and F1 value are all improved for the Shanghai Stock Connect Index (000159). F1 value is increased from 0.740741 to 0,888889. The result indicates that the ability of the new model in finding out the correct up directions and avoid the incorrect classification of stock price movement directions is much better than the normal nonlinear autoregressive model without the exogenous sentiment inputs.

To visualize the performance of the model, Fig. 6 is presented. Figure 6 compares the prediction output of the proposed model and the expected or target value for the stock 388. According to the figure, the two time series, the output price and the target price, have almost the same curve and moving directions. The error or the difference between the output price and the target price is quite small which demonstrates the prediction ability of the proposed model. Most of the error locate in the range -2 to 2, only a few exceed the range and the largest error is around 5 only.

Table 6
Predicting the positive moving direction for 000159

	I_1	I_0
Precision	0.923077	0.769231
Recall	0.857143	0.714286
F1 value	0.888889	0.740741

8. Conclusion

As the stock market in China has already been one of the largest market in the world, it is important to understand its behavior according to the public mood states. This paper investigates whether public mood states derived from large-scale collection of Weibo posts on weibo.com is correlated or predictive of the price of different stocks in China market. The results demonstrate the public mood states can be indeed tracked using the proposed new Chinese sentiment analysis method on the large-scale Weibo posts.

In addition, the Granger causality analysis results indicate that the null hypothesis that the public mood is not correlated with the stock price can be rejected and the time series of public mood states is predictive of the price movement of different stocks. Among the 7 observed emotions, only some are Granger causative of the stock price movement. The emotions "Happiness" and "Disgust" have significant causative relationships with different stock prices. Movement of these two emotions is correlated with the price movement which occurs in 3–4 days. These 3 days time lag is also discussed in the research of using public mood in Twitter to predict stock price [3]. A nonlinear autoregressive model with exogenous sentiment inputs furthermore demonstrates the possibility of using public mood states to predict the stock price movement in China. The public mood states help to improve the prediction accuracy even using the basic neural network model. Given the performance, if more related topics for a stock are found out, the accuracy could be higher considering more completed topic-based sentiment inputs. Regarding the model, if more sophisticated and complicated model is implemented according to the topic-based mood, it is possible that the prediction result would be even better.

In the future, more experiments will be conducted to understand why the time lag exists and how to avoid and use it for real time algorithm trading. Other important factors will be examined in future research such as the geo-location effect and spam detection. In addition, in this paper, it only discusses the relationship between the close price of stocks and the daily movement of public mood. How the intraday public mood change affects the real-time stock price movement will be discussed in future research as well.

Acknowledgement

This paper is an extended version from the conference paper "Weibo Mood towards Stock Market" in SeCoP 2016. Supported by Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (No. 2015TQ01X633), (STPP-GD) (Grant No. 2015A070711001) and (SF-STRDA-GD) (Grant No. 2016B010124011).

References

[1] A. Abbasi and H. Chen, Cybergate: A design framework and system for text analysis of computer-mediated communication, *MIS Quarterly* (2008), 811–837.

[2] K. Ahmad, D. Cheng and Y. Almas, Multi-lingual sentiment analysis of financial news streams, in: *Proceedings of the 1st International Conference on Grid in Finance*, 2006.

[3] J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, *Computational Science* **2**(1) (2011), 1–8. doi:10.1016/j.jocs.2010.12.007.

[4] W. Brock, J. Lakonishok and B. LeBaron, Simple technical trading rules and the stochastic properties of stock returns, *The Journal of Finance* **47**(5) (1992), 1731–1764. doi:10.1111/j.1540-6261.1992.tb04681.x.

[5] M. Chau and J. Xu, Mining communities and their relationships in blogs: A study of hate groups, *International Journal of Human-Computer Studies* **65**(1) (2007), 57–70. doi:10.1016/j.ijhcs.2006.08.009.

[6] J.M. Chen, The construction and application of Chinese emotion word ontology, Master thesis, Dailian University of Technology, 2008.

[7] R.J. Dolan, Emotion, cognition, and behavior, *Science* **298**(5596) (2002), 1191–1194. doi:10.1126/science.1076358.

[8] E.F. Fama, The behavior of stock-market prices, *The Journal of Business* **38**(1) (1965), 34–105. doi:10.1086/294743.

[9] E. Fama, The behavior of stock-market prices, *Business* **38**(1) (1965), 34–105. doi:10.1086/294743.

[10] E.F. Fama, Efficient capital markets: II, *The Journal of Finance* **46**(5) (1991), 1575–1617. doi:10.1111/j.1540-6261.1991.tb04636.x.

[11] R. Fan, J. Zhao, Y. Chen and K. Xu, Anger is more influential than joy: Sentiment correlation in weibo, *PloS One* **9**(10) (2014), e110184. doi:10.1371/journal.pone.0110184.

[12] Q. Gao, F. Abel, G.J. Houben and Y. Yu, A comparative study of users' microblogging behavior on Sina Weibo and Twitter, in: *User Modeling, Adaptation, and Personalization*, Springer, 2012, pp. 88–101. doi:10.1007/978-3-642-31454-4_8.

[13] E. Gilbert and E. Karahalio, Widespread worry and the stock market, in: *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010, pp. 59–65.

[14] D. Gruhl, R. Guha, R. Kumar, J. Novak and A. Tomkins, The predictive power of online chatter, in: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 78–87. doi:10.1145/1081870.1081883.

[15] A. Gunasekarage and D.M. Power, The profitability of moving average trading rules in South Asian stock markets, *Emerging Markets Review* **2**(1) (2001), 17–33. doi:10.1016/S1566-0141(00)00017-0.

[16] K. Huarng and T.H.-K. Yu, The application of neural networks to forecast fuzzy time series, *Physica A: Statistical Mechanics and Its Applications* **363**(2) (2006), 481–491. doi:10.1016/j.physa.2005.08.014.

[17] T. Kimoto, K. Asakawa, M. Yoda and M. Takeoka, Stock market prediction system with modular neural networks, in: *Proceedings of the International Joint Conference on Neural Networks*, 1990, pp. 1–6.

[18] A. Lapedes and R. Farber, Nonlinear signal processing using neural networks: Prediction and system modelling, Technical Report LA–UR–7–2662.

[19] G. Leng, G. Prasad and T.M. McGinnity, An on-line algorithm for creating self-organizing fuzzy neural networks, *Neural Networks* **17**(10) (2004), 1477–1493. doi:10.1016/j.neunet.2004.07.009.

[20] X. Li, H. Xie, L. Chen, J. Wang and X. Deng, News impact on stock price return via sentiment analysis, *Knowledge-Based Systems* **69** (2014), 14–23. doi:10.1016/j.knosys.2014.04.022.

[21] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li and F.L. Wang, Does summarization help stock prediction? A news impact analysis, *IEEE Intelligent Systems* **30**(3) (2015), 26–34. doi:10.1109/MIS.2015.1.

[22] H. Liang, F.S. Tsai and A.T. Kwee, Detecting novel business blogs, in: *Proceedings of the 7th International Conference on Information*, 2009, pp. 1–5.

[23] A.Y. Liu, B. Gu, P. Konana and J. Ghosh, Predicting stock price from financial message boards with a mixture of experts framework, in: *Intelligent Data Exploration & Analysis Laboratory*, 2006, pp. 1–14.

[24] G. Mishne and N. Glance, Predicting movie sales from blogger sentiment, in: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 155–158.

[25] C.R. Nelson and C.R. Plosser, Trends and random walks in macroeconmic time series: Some evidence and implications, *Monetary Economics* **10** (1982), 139–162. doi:10.1016/0304-3932(82)90012-5.

[26] J.R. Nofsinger, Social mood and financial economics, *Behaviour Finance* **6**(3) (2005), 144–160. doi:10.1207/s15427579jpfm0603_4.

[27] D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", *Decision Support Systems* **51**(4) (2011), 821–830. doi:10.1016/j.dss.2011.01.016.

[28] P.F. Pai and C.S. Lin, A hybrid arima and support vector machines model in stock price forecasting, *Omega* **33**(6) (2005), 497–505. doi:10.1016/j.omega.2004.07.024.

[29] B. Qian and K. Rasheed, Stock market prediction with multiple classifiers, *Applied Intelligence* **26**(1) (2007), 25–33. doi:10.1007/s10489-006-0001-7.

[30] Y. Rao, H. Xie, J. Li, F. Jin, F.L. Wang and Q. Li, Social emotion classification of short text via topic-level maximum entropy model, *Information & Management* (2016).

[31] R.P. Schumaker and H. Chen, Textual analysis of stock market prediction using breaking financial news: The azfintext system, *ACM Transactions on Information System* **27**(2) (2009), 1–19. doi:10.1145/1462198.1462204.

[32] J. Sun, G. Wang, X. Cheng and Y. Fu, Mining affective text to improve social media item recommendation, *Information Processing and Management* **51** (2015), 444–457. doi:10.1016/j.ipm.2014.09.002.

[33] M.H. Wang and C.L. Lei, Boosting election prediction accuracy by crowd wisdom on social forums, in: *13th IEEE Annual Consumer Communications & Networking Conference*, 2016, pp. 348–353.

[34] F. Yang, Y. Liu, X. Yu and M. Yang, Automatic detection of rumor on Sina Weibo, in: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, 2012, p. 13.

[35] X. Zhang, H. Fuehres and P.A. Gloor, Predicting stock market indicator through Twitter "I hope it is not as bad as I fear", *Procedia-Social and Behavioral Sciences* **26** (2011), 55–62. doi:10.1016/j.sbspro.2011.10.562.

[36] X. Zhu, H. Wang, L. Xu and H. Li, Predicting stock index increments by neural networks: The role of trading volume under different horizons, *Expert Systems with Applications* **34**(4) (2008), 3043–3054. doi:10.1016/j.eswa.2007.06.023.