Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis

Yifan Liu[1], Zengchang Qin*[1], Pengyu Li[1,2], and Tao Wan*[3]

[1]Intelligent Computing and Machine Learning Lab, School of ASEE
Beihang University, Beijing 100191, China
[2]School of Mechanical Engineering and Automation
Beihang University, Beijing, 100191, China
[3]School of Biological Science and Medical Engineering
Beihang University, Beijing, 100191, China
*{taowan,zcqin}@buaa.edu.cn

Abstract. In this paper, we propose a model to analyze sentiment of online stock forum and use the information to predict the stock volatility in the Chinese market. We have labeled the sentiment of the online financial posts and make the dataset public available for research. By generating a sentimental dictionary based on financial terms, we develop a model to compute the sentimental score of each online post related to a particular stock. Such sentimental information is represented by two sentiment indicators, which are fused to market data for stock volatility prediction by using the Recurrent Neural Networks (RNNs). Empirical study shows that, comparing to using RNN only, the model performs significantly better with sentimental indicators.

Keywords: Natural language processing, Stock volatility prediction, Sentimental analysis, Sentimental score

1 Introduction

In time-series data mining, stock market is notoriously difficult to analyze, even be totally unpredictable based on the famous Efficient Market Hypothesis (EMH). As early as 1900s, Bachelier [2] applied statistical methods to analyze stock data, and found that the mathematical expectation of the stock fluctuation tends to be zero. In the 1970s, Fama [7] formally put forward the EMH, which stated that under the condition of market with complete information, investors couldn't gain more than fifty percent of the profits only with the past price, or simply, no one could 'beat' the market' continuously. The Random Walk Theory (RWT) proposed by Osborne [14] also suggested same conclusion that the stock prices were unpredictable. But all these theories are based on the same assumption that investors are rational and complete market information is available.

Paul Hawtin[1] once said "For years, investors have widely accepted that financial markets are driven by fear and greed." In the actual market, investors

[1] The founder of Derwent Capital Markets and one of early pioneers in the use of social media sentiment analysis to trade financial derivatives.

cannot be completely rational. They may be influenced by their emotions and make impulsive decisions [16]. Therefore, the basic assumption of EMH and RWT is not impregnable. Many researchers have tried to study the correlation between sentiment and stock market volatility to challenge the classical theories. For example, researchers found that some factors, e.g. weather and sports games, can affect public emotion and also the stock market. The sunny weather and the rising stock index had certain correlations [10]. There would be a significant market decline after the soccer lost [5]. In recent years, the rapid development of social networks (Facebook, Twitter, Weibo) opens a new door to measure the public emotion. Bollen et al. [3] analyzed the text content of daily Twitter by using two mood tracking tools: OpinionFinder (OF) and Google-Profile of Mood States (GPOMS). The authors then used self-organizing fuzzy neural network (SOFNN) to predict the volatility of the Dow Jones Industrial Average (DJIA). By considering the sentimental information from Twitter, the prediction accuracy has raised up by 13%. The results were very encouraging and this direction was followed by some other similar research. Zhang et al. [19] found that a burst of public emotion no matter positive or negative, heralded the falling of the index. There were also some research on the individual stock, Si et al. [17] proposed a technique to leverage topic based sentiment from Twitter to help predict the stock price while O'Connor [13] found that the popularity of a brand was much related to related tweets and its stock price.

But there are still some problems remained. First, Twitter users are predominantly English speakers, or even worse the investors of a particular market may not use Twitter to discuss their finance [3]. Second, popular sentiment analysis dictionary can not entirely measure the emotion of the stock investors [11]. Sprenger et al. [18] selected tweets which mentioned the company in the Standard & Poor's 100 index, and labeled the tweets with *buy*, *hold* or *sell* signals. With the labeled training data, they used a Naive Bayes classifier to extract the signals from the tweets automatically and calculated the bullishness through these signals. Finally, they found that a strategy based on bullishness signals could earn substantial abnormal returns.

Above all, a great amount of focus has been placed on the correlation between the investors' sentiment and the U.S. stock market. While limited by the Chinese expression complexity, little attention has been paid to the the relevant research on Chinese stock market. According to the World Federation of Exchanges database[2], Chinese market capitalization ranked the second in 2015 in the world. That's the reason we focus on our study of Chinese stock market. In this paper, based on Sprenger's [18] approach, we propose a model to study Chinese stock market. The sentiment of Chinese investors are from the East Money Forum[3], which is one of the biggest and specified stock forum in China, but it is not public forums like Facebook or Twitter. Each stock has its individual sub-forum which ensures that most posts from the sub-forums are published by the investors who hold or sell this particular stock. In order to

[2] http://www.indexmundi.com/facts/indicators/CM.MKT.LCAP.CD/rankings

[3] http://guba.eastmoney.com/

avoid the problem that the ordinary dictionary often makes misunderstanding in recognizing investors sentiment, we use a machine learning method to generate our own dictionary and then calculate the sentiment score of the posts based on the dictionary automatically. To study the correlation between Chinese stock market and Chinese investor sentiment, we propose sentiment indicators for the stock volatility prediction model using the Recurrent Neural Networks (RNNs) to obtain a better performance.

2 Sentiment Analysis

Bollen et al. [3] proposed a dictionary-based method for sentiment analysis of the financial contexts. However, Loughran and Mcdonald [11] found that three-fourths of the words identified as negative by the Harvard Dictionary are not typically considered as negative in financial contexts. The same problem also occurs in Chinese sentiment analysis. We find that Chinese posts in stock forums have some special expressions containing strong emotions. But these expressions rarely appear in common sentiment analysis dictionary. So in this research, we first need a practical dataset from which we can obtain a dictionary of financial words, we also develop a simple but effective tool to generate sentimental weights for the words.

2.1 Data Processing

East Money Forum is one of the most influential Internet financial media in China. It has more than 3000 sub-forums for each individual stock. We randomly select 10 stocks as well as the sub-forum of the posts from 25^{th} Sept., 2015 to 30^{th} Sept., 2016 with a web crawler "Bazhuayu (means "Octopus")"[4]. Nearly 96000 pieces of posts are obtained, and most of them are short and colloquial. They do not follow any strict syntax but contain strong sentiment. We randomly sampled 3427 stock posts from 10 different stocks to do the manual annotation[5]. If the post expresses an optimistic attitude towards the stock market and suggests to buy, we label it as positive, otherwise, we label it as negative. We have annotated 2067 negative posts and 1360 positive posts manually. The original Chinese texts need to be preprocessed by segmentation, and a classical Chinese text segmentation tool called "Jieba" (Chinese for "to stutter") in Python[6] is chosen for this.

2.2 Polarity Model of Sentiment

The polarity model of sentiment is trained by a collection of texts labelled only by positive or negative. Emotional words are extracted and each one has an

[4] http://www.bazhuayu.com

[5] http://dsd.future-lab.cn/members/2016/LiuYFProject/data.xlsx

[6] https://github.com/fxsjy/jieba

associated sentiment weight. Weights can be learned from the labeled training dataset [9]. The sum of weighted sentiment scores of all terms determines the sentiment polarity (positive or negative) of the post. If the sum is greater than 0, it is positive and vice versa. The sentiment score $h_{\mathbf{w}}(\mathbf{x})$ of a given post is computed as follows:

$$h_{\mathbf{w}}(\mathbf{x}) = f\left(\sum_{i=1}^{N} w^{(i)} x^{(i)}\right) = f(\mathbf{w}^T\mathbf{x}); 1 \leq i \leq N \tag{1}$$

where N is the number of all the terms in the corpus, a term could be uni-gram or bi-gram model. $w^{(i)}$ is the sentimental weight for each term $t^{(i)}$, $x^{(i)}$ is the term frequency or tf-idf value of the given term $t^{(i)}$. Function $f(\cdot)$ is a sigmoid function to compress the linear combination of sentimental weight into 0 and 1, and make it smooth.

$$f(z) = \frac{1}{1+e^{-z}} \tag{2}$$

By using the logistic regression, the target label y is 1 for positive posts and 0 for negative ones. So that $h_{\mathbf{w}}(\mathbf{x})$ represents the probability of a post being positive. If we take the threshold value as 0.5, the prediction of the sentiment is:

$$y = \begin{cases} 1 \; h > 0.5 \\ 0 \; h \leq 0.5 \end{cases} \tag{3}$$

Given a training corpus with M text posts, $\mathbf{x}^{(k)}$ denotes the k^{th} $(1 \leq k \leq M)$ post feature value vector. We can derive the cost function and its logarithmic likelihood function based on the maximum likelihood estimation. The loss function $J(\cdot)$ is:

$$J(h_{\mathbf{w}}(\mathbf{x}), y) = \begin{cases} -\log(h_{\mathbf{w}}(\mathbf{x})) & if\ y = 1 \\ -\log(1 - h_{\mathbf{w}}(\mathbf{x})) & if\ y = 0 \end{cases} \tag{4}$$

The average loss for the entire data set is (for $1 \leq k \leq M$):

$$\begin{aligned} J(\mathbf{w}) &= -\frac{1}{M}\sum_{k=1}^{M} J(h_{\mathbf{w}}(\mathbf{x}), y) \\ &= -\frac{1}{M}\left[\sum_{k=1}^{M} y^{(k)} \log(h_{\mathbf{w}}(\mathbf{x}^{(k)})) + (1-y^{(k)})\log(1-h_{\mathbf{w}}(\mathbf{x}^{(k)}))\right] \end{aligned} \tag{5}$$

In order to minimize $J(\mathbf{w})$, we can update $\mathbf{w}$ using the Gradient Descent algorithm with a learning rate α: $\mathbf{w}_{j+1}^{(k)} = \mathbf{w}_j^{(k)} - \alpha(h^{(k)} - y^{(k)})\mathbf{x}^{(k)}$. The values of the sentimental weight can be obtained. The term with a higher sentimental weight indicates a stronger positive sentiment and vice versa. We use the weights and the corresponding terms to build a sentimental dictionary. The sentiment score of a post can be calculated by weighted sum of weights of all consisting terms based on Eq.(1).

3 Emotion Model for Stock Prediction

3.1 Sentimental Indicators

Some literatures in finance [18] suggest that individual investors have a herd mentality when they make decisions. For example, if they find that most of people are not optimistic in the outlook of the stock price, they will trade on the advice and move the price. What's more, a larger quantity of the posts on a forum indicates a larger amount of attention which may lead to a severe price volatility. Therefore, we propose an emotion model (EMM) according to the following two important assumptions: (1) Increased bullishness of stock posts is associated with higher stock price. (2) Increased posts volume suggests a more substantial volatility. The index of bullishness of online posts can be defined on a daily basis according to [1]:

$$B_t = \ln \frac{1 + N_t^p}{1 + N_t^n} \tag{6}$$

where $N_t^p (N_t^n)$ represents the number of positive (negative) posts on the day t. This indicator reflects both the expectations of the rise in price and the total number of posts. When the posts have a continuous sentimental score instead of a binary label, the index of bullishness becomes

$$B_t = \ln \frac{\varepsilon + S_t^p}{\varepsilon + |S_t^n|} = \ln \frac{\varepsilon + \sum_{k=1}^{N_t^p} h_{\mathbf{w}}(\mathbf{x})^{(k)}}{\varepsilon + \left| \sum_{i=1}^{N_t^n} h_{\mathbf{w}}(\mathbf{x})^{(i)} \right|} \tag{7}$$

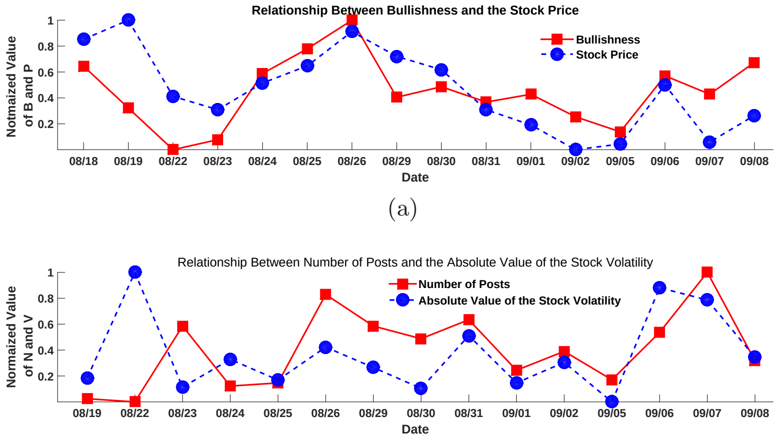
where $S_t^p (S_t^n)$ represents the sum of positive (negative) sentimental score of the posts on the day t and $h_{\mathbf{w}}(\mathbf{x})^{(k)}$ $(h_{\mathbf{w}}(\mathbf{x})^{(i)})$ represents the positive (negative) sentimental score of the k^{th} (i^{th}) post on the day t. $\varepsilon (\varepsilon > 0)$ is a tiny number for smoothing, and we set $\varepsilon = 0.0001$ in our research. The reason we use absolute value is because the score of negative sentiment is always less than zero.

The total number of the posts N_t on the day t is $N_t = N_t^p + N_t^n$. To enable fair comparison for B_t and N_t, we use the z-score to normalize data based on the mean and standard deviation within a sliding window of length l (we average the data of l days before and after the current date t). The z-scores for B_t and N_t are:

$$Z_B{}^{(t)} = \frac{B_t - \mu(B_{t \pm l})}{\sigma(B_{t \pm l})} \tag{8}$$

$$Z_N{}^{(t)} = \frac{N_t - \mu(N_{t \pm l})}{\sigma(N_{t \pm l})} \tag{9}$$

where $\mu(B_{t \pm l})$ and $\mu(N_{t \pm l})$ are the means and $\sigma(B_{t \pm l})$ and $\sigma(N_{t \pm l})$ are the standard deviations with $2l$ days around the current day t. The correlations of Z_B and stock price (Fig. 1-(a)), Z_N and stock volatility (Fig. 1-(b)) given a particular stock are shown in Fig. 1 We can see they are positively correlated and satisfy the two assumptions on sentimental indicators we previously gave.



(a)

(b)

Fig. 1: Relationship between sentimental indicators and the stock information

3.2 Stock Prediction with Recurrent Neural Network

The stock price for a day is the weighted average price of all transactions on that day. In the Chinese market, it is calculated by the last minute of the trading day, so it is also referred to as the closing price [12]. In the actual stock market, profit-driving investors only care about the volatility of a stock instead of the exact price. The stock volatility V_t is defined based on the closing price P_t on the current day t and the previous day $t-1$:

$$V_t = \frac{P_t - P_{t-1}}{P_{t-1}}; V_t \in [-0.1, 0.1] \tag{10}$$

In order to regulate the stock market from any malicious manipulations, any stock has the volatility more than 10% will be forced to quit the market on that trading day, therefore, V_t always lies in the range of $[-0.1, 0.1]$. In our experiment, we normalize the volatility into a time series between 0 and 1 based on mini-max normalization method, and generate the normalized time series $\mathbf{V}$. At the same time, we set 0.5 as the threshold in order to obtain a binary label (0 for price going down, and 1 for price rising).

$$F_t = \begin{cases} 1 & V_t > 0.5 \\ 0 & otherwise \end{cases} \tag{11}$$

A stock market is highly complex in the control of “invisible hands”. However, there are still loads of research on statistical modeling and machine learning approaches to learn from history data. The key to predict the stock market is to fit a latent nonlinear relation between the history data and the future stock volatility. The traditional statistical models used for financial forecasting were

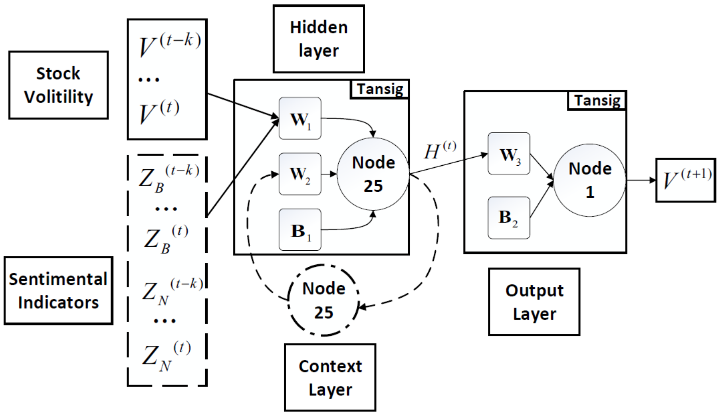


Fig. 2: The structure of the RNN model with sentimental indicators. The input values are stock volatility (V) and sentimental indicators (Z), we use the inputs of previous k trading days to predict the stock volatility of the next trading day ($V^{(t+1)}$). There are 25 hidden nodes used in our model.

simple and suffered from several shortcomings. Machine learning methods like Multi-Layer Perception (MLP), Recurrent Neural Networks, Support Vector Machine (SVM) [15] have an increasing popularity in this area.RNN is incorporated in our fundamental prediction model due to its appropriateness to address time series problem. The context layer stores the outputs of the state neurons from the previous time step and outputs to the next time step for computation. In this paper, we employ Elman Network [6] in the following experiments. If we denote the output of hidden layer at time t by $H^{(t)}$, the final prediction can be made by:

$$V^{(t+1)} = f(H^{(t)}W_3 + B_2) \tag{12}$$

$$H^{(t)} = f([V, Z]W_1 + H^{(t-1)}W_2 + B_1) \tag{13}$$

where $f(\cdot)$ is the activation function and B is bias. The structure of our proposed model [7] is show in Fig. 2.

4 Experimental Studies

In order to verify the effectiveness of the new proposed model, we test it on the stock data introduced in Section 2.1. The stock data of 250 consecutive trading days is downloaded from the DaZhiHui (DZH)[8] software. In order to evaluate

[7] github link: `https://github.com/irfanICMLL/EMM-for-stock-prediction`.

[8] It can be downloaded from `http://www.gw.com.cn`.

the quality of the model, we define the concept of accuracy based on the binary label F. F^* is the predict label of the test data while F is the real label. Define $counter$ as the total number if ${F^*}_t = F_t$, accuracy $Acc = \frac{counter}{\|F\|}$.

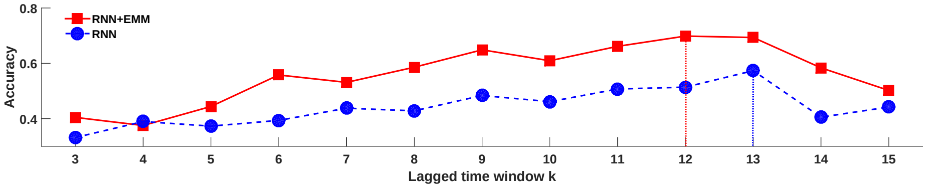


Fig. 3: Comparison results with different k.

We choose one stock (000573) as an example, we extract the volatility data and run RNN on it, and then compare to the RNN with sentimental indicators ${Z_B}^{(t)}$ and ${Z_N}^{(t)}$ (RNN+EMM). We vary k from 3 to 15 to test the best length of history for predicting the future. The experiments are replicated for 50 times. Comparison results are shown in Fig. 3. We can see that sentimental indicators help to improve the accuracy significantly, and the parameter k will affect the prediction accuracy, the optimal length is around 10 based on different data sets. For the stock 000573, the best accuracy of the EMM with RNN is 69.85%$(k = 12)$ while the best accuracy without sentimental information is 57.33%$(k = 13)$, and the accuracy is significantly better than 0.5. Another 9 stocks are selected randomly to test the model, that increase the credibility of the conclusion.

As for each particular stock, we can obtain better performance for 8 datasets in 10 and the detailed result comparisons are shown in Table. 1 . To make it more intuitive, we draw the histogram in Fig. 4. From the results, we can see that the stock 000573 performs better than others. The reason may be that most of the training posts of the emotion classifier come from its sub-forum during the chosen period. In other words, if the actual sentimental indicators are obtained, the accuracy of the model can be better.

Table 1: Accuracy and the best k for RNN+EMM and RNN

stock number	000573	000733	000703	300017	600605	300333	000909	601668	000788	600362
RNN+EMM	0.6985	0.6187	0.6757	0.6927	0.7355	0.6491	0.6154	0.5626	0.5917	0.7092
RNN	0.5733	0.524	0.605	0.6455	0.6982	0.6232	0.6029	0.5543	0.6017	0.7344
k (RNN+EMM)	12	13	11	14	3	4	11	15	13	11
k (RNN)	13	5	5	14	14	6	10	3	13	6

Table 2 shows the comparison results of four learning models: MLP [8], SVM [4], RNN, EMM+RNN and the baseline is a random guesser (RAND). On the 10 datasets with online discussions, the accuracy of RNN is higher than MLP

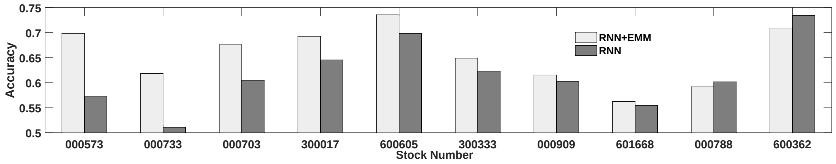


Fig. 4: Accuracy for RNN+EMM and RNN on 10-stocks dataset

and SVM, because it contains information about the previous states. When considering sentimental indicators, the prediction performance improves nearly 4% on the average which verifies the assumptions about the sentimental indicators. We only test 10 stocks, because the posts are not so easy to be obtained and labeled. We need roughly 20 hours to collect one sub-forum. And then extract the sentimental indicators through the method introduced in Section 2. In order to make the results more credible, we do repetition under various of the initial parameters to make sure that the improved accuracy is thus most likely not the result by chance nor the selection of a specifically favorable test period.

Table 2: Performance comparisons on 10 stocks from the Chinese market.

Method	RAND	MLP	SVM	RNN	RNN+EMM
MEAN	0.500168	0.559257	0.602339	0.61623	0.6549
STD	0.003846	0.028681	0.111318	0.06347	0.05640

5 Conclusions

In this research, we investigated the relationship between the stock volatility and sentimental information obtained from an online stock forum. We employed a RNN model to consider sentimental information, experimental results show that the new model can boost the prediction accuracy. The main contribution of our research are as follows: (1) Generate a sentimental weight dictionary of Chinese stock posts. (2) Propose sentimental indicators and investigate the relationship between the stock volatility and the information from the stock forums. (3) Build a RNN model considering sentimental information for stock prediction and verifies the information from forums can help to predict the stock market of China. (4) We construct a benchmark dataset of labeled financial posts and make it public available for comparison studies. Finally, it's worth mentioning that our analysis doesn't take into account many factors. The posts from the forums may contains a lot of fake messages that confuse the public. We will consider that in our future work.

Acknowledgement

This work is supported by the National Science Foundation of China Nos. 61401012 and 61305047.

References

1. Antweiler, W., Frank, M.Z.: Is all that talk just noise? the information content of internet stock message boards. Journal of Finance 59(3), 1259–1294 (2004)
2. Bachelier, L.: Théorie de la spéculation. Annales Scientifiques De L École Normale Supérieure 3, 21–86 (1900)
3. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2010)
4. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
5. Edmans, A., García, D., Norli, Ø.: Sports sentiment and stock returns. Journal of Finance 62(4), 1967–1998 (2007)
6. Elman, J.L.: Finding structure in time. Cognitive Science 14(2), 179–211 (1990)
7. Fama, E.: Efficient market hypothesis: A review of theory and empirical work. Journal of Finance 25, 383–417 (1970)
8. Frank, R.J., Davey, N., Hunt, S.P.: Input window size and neural network predictors. In: IEEE-INNS-ENNS International Joint Conference on Neural Networks. vol. 2, pp. 237–242 (2000)
9. Guo, T., Li, B., Fu, Z., Wan, T., Qin, Z.: Learning sentimental weights of mixed-gram terms for classification and visualization. PRICAI 2016, LNAI 9810 pp. 116–124 (2016)
10. Hirshleifer, D., Shumway, T.: Good day sunshine: Stock returns and the weather. Journal of Finance 58(3), 1009–1032 (2003)
11. Loughran, T., Mcdonald, B.: When is a liability not a liability? CFA Digest 41(2), 57–59 (2011)
12. Ma'Aji, M.M., Abdullahi, S.R.: Market reaction to international cross-listing: Evidence from nigeria. Social Science Electronic Publishing (1), 13–25 (2014)
13. O'Connor, A.: The power of popularity: An empirical study of the relationship between social media fan counts and brand company stock prices. Social Science Computer Review 31(2), 229–235 (2013)
14. Osborne, M.: Browing motion in the stock market. Operations Research 7(2), 145–173 (1959)
15. Rout, A.K., Dash, P.K., Dash, R., Bisoi, R.: Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach. Journal of King Saud University - Computer and Information Sciences (2015)
16. Rzepczynski, M.: Beyond greed and fear: Understanding behavioral finance and the psychology of investing. OUP Catalogue (78), 99–101 (2007)
17. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., Deng, X.: Exploiting topic based twitter sentiment for stock prediction. Proceedings of ACL pp. 24–29 (2013)
18. Sprenger, T.O., Tumasjan, A., Sandner, P.G., Welpe, I.M.: Tweets and trades: the information content of stock microblogs. European Financial Management 20(5), 926–957 (2010)
19. Zhang, X., Fuehres, H., Gloor, P.A., Zhang, X., Fuehres, H.: Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Social and Behavioral Sciences 26(26), 55–62 (2011)