Can Online Emotions Predict the Stock Market in China?

Zhenkun Zhou[1], Jichang Zhao[2(✉)], and Ke Xu[1]

[1] State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2] School of Economics and Management, Beihang University, Beijing, China
jichang@buaa.edu.cn

Abstract. Whether the online social media, like Twitter or its variant Weibo, can be a convincing proxy to predict the stock market has been debated for years, especially for China. However, as the traditional theory in behavioral finance states, the individual emotions can influence decision-makings of investors, so it is reasonable to further explore this controversial topic from the perspective of online emotions, which is richly carried by massive tweets in social media. Surprisingly, through thorough study on over 10 million stock-relevant tweets from Weibo, both correlation analysis and causality test demonstrate that five attributes of the stock market in China can be competently predicted by various online emotions, like disgust, joy, sadness and fear. Specifically, the presented model significantly outperforms the baseline solutions on predicting five attributes of the stock market under the K-means discretization. We also employ this model in the scenario of realistic online application and its performance is further testified.

Keywords: Social media · Stock market · Sentiment analysis · Causality test

1 Introduction

With explosive development of online social media, tremendous amounts of tweets are posted and reposted in popular platforms like Twitter and Weibo. These tweets, spreading in terms of word-of-mouth, not only convey the factual information, but also reflect the emotional statuses of the authors. Taking Weibo as an example, around 100 million Chinese tweets are posted every day and from which we can not only sense what happens in China, but also how 500 million users feel about their lives. In fact, the online social media indeed provide us an unprecedented opportunity to study the detailed human behavior from many new views. The investment decision in the stock market, as one of the most important issues, attracts much attention in recent decades.

However, whether online social media like Twitter can be excellent predictors is still controversial, especially for the stock market in China [2,9,16]. Different from the west, the marketing policy intervention in China will introduce

W. Cellary et al. (Eds.): WISE 2016, Part I, LNCS 10041, pp. 328–342, 2016.
DOI: 10.1007/978-3-319-48740-3_24

more non-market factors that might disturb the fluctuation of the stock market. And moreover, those possible interventions could be leaked through the social media and then greatly influence the investors' emotions and decisions. In the mean time, considering the irrationality of huge amount of individual investors in China (which is also rare in the west), their actions might be more easily affected by online news and other investors' feelings about the market. Then the messages about the stock market and the sentiments they convey could be good indicators for the market prediction. Thus, like the conventional behavioral finance theory claims, which the emotion can influence the decision-process of the investors, it is necessary to investigate the following important issues:

- Is there indeed significant correlation between online emotions and attributes of Chinese stock market?
- Can online emotions predict the attributes of the stock market in China?
- Which emotion does play the critical role in predicting various attributes of the Chinese stock market?

In the present study, we collect over 10 million Chinese stock-relevant tweets from Weibo and classify them into five emotions, including anger, disgust, joy, sadness and fear. Besides the daily closing index of Shanghai Stock Exchange[1], we consider the daily opening index, the intra-day highest index, the intra-day lowest index and the daily trading volume of the stock market. By both correlation analysis and Granger causality test, it is revealed that disgust has a Granger causal relation with the closing index, joy, fear and disgust have Granger causal relations with the opening index, joy, sadness and disgust have Granger causal relations with the intra-day highest and lowest index, and correlation between trading volume and sadness is unexpectedly strong. It's also surprising to find that anger in online social media possesses the weakest correlation or even is no relation with the Chinese stock market.

Based on the findings, we develop classification-oriented predictors, in which different emotions are selected as features, to predict five daily attributes of the stock market in China. The comparison with other baseline methods show that our model can outperform them according to K-means discretization. And the model is also deployed in a realistic application and achieves the accuracy of 64.15 % for the intra-day highest index (3-categories) and the accuracy of 60.38 % for the trading volume (3-categories). Our explorations demonstrate that the online emotions, specially disgust, joy, sadness and fear, in Weibo indeed can predict the stock market in China.

2 Related Works

Emotion expression and stock fluctuation are usually bonded together in the traditional theory and even in social media. Behavioral economics studies the

[1] In the paper, index refers in particular to Shanghai Stock Exchange Composite Index and the trading volume refers in particular to the daily volume of the Shanghai Stock Exchange.

effects of social, emotional and psychological factors on the economic decisions of individuals and institutions and the consequences for market prices. It demonstrates that mood can affect individual behavior and decision-makings of investors [6,10].

Owing to lack of effective measurement method of sentiment, stock prediction using emotions had been in dispute [1,4]. However, with the recent widespread presence of computers and Internet, public emotions can be extracted from data on online platforms. Using Twitter as a corpus, some researchers built sentiment classifiers, which are able to determine different sentiments for a tweet [13] [12] [8]. Specially on Sina Weibo platform, Zhao et al. trained a fast Naive Bayes classifier for Chinese emotion classification, which is now available online for temporal and spatial sentiment pattern discovery [18].

In addition, there have long been controversies on predictive power of social media aiming at different fields [7,14]. In the field of finance, Bollen et al. found that public mood on Twitter could predict the Dow Jones Industrial Average [2]. The public mood dimensions of Calm and Happiness seemed to have a predictive effect. However, the tweets they collected were associated with whole social status, not just the stock market in America, which could not represent online investors' sentiment. Oh et al. also showed stock micro-blog sentiments did have predictive power for market-adjusted returns. Instead of emotion on social media [11], some researchers examined textual representations in financial news articles for stock prediction [5,15]. Ding et al. proposed a deep learning method for event-driven stock market prediction on large-scale financial news dataset [17]. Besides, Bordino et al. showed that daily trading volumes of stocks traded in NASDAQ100 were correlated with daily volumes of queries related to the same stocks [3].

However, to the best of our knowledge, existing studies referring to the stock market in China are relatively few. Mao et al. pointed out that Twitter did not have a predictive effect, as regards to predicting developments in Chinese stock markets [9]. They advised adopting the tweets on Weibo platform to research Chinese stock market. Based on 66,317 tweets of Weibo with one year and two emotion categories, Cheng and Lin found that the investors' bullish sentiment of social media can help to predict trading volume of the stock market, but still does not work for the market returns [16]. Because of less collected data set and simple emotion classification, it is not easy to generalize their conclusions to other realistic scenarios.

While in this paper, we focus purely on the stock market in China and try to understand the predictive ability of multiple online emotions in Weibo. Different from the previous study, we hope to develop predictors from more data sets and more sentiments and to predict more attributes of the real market.

3 Data Sets

3.1 Online Stock Market Emotions

The feelings of investors can be collected through many different approaches, like questionnaires in previous study. While with the explosive development of the social media in China, more and more investors express their seeings, hearings and feelings on Weibo. Therefore, we choose and utilize the characteristic of Weibo to obtain online emotion referring to Chinese stock market.

From December 1st 2014 to December 7th 2015, the massive public tweets on Weibo are collected through its open APIs. However, only a fraction of the tweets are semantically related with Chinese stock market. Filtering out the irrelevant tweets and remaining the data that truly represents the stock market emotion is a very significant step. Therefore, we manually select six Chinese keywords, including Stock, Stock Market, Security, The Shenzhen Composite Index, The Shanghai Composite Index and Component Index with help of expertise from the background of finance. These manually selected keywords are supposed to depict the overall status of Chinese stock market sufficiently. We postulate that if the text of tweet contains one or more of the six selected keywords, the tweet describes the news, opinions or sentiments about Chinese stock market. In our database, the number of tweets related to stock market, involving one or more keywords, is a total of 10,550,525 from December 1st 2014 to December 7th 2015.

In the paper, the emotions are divided into five categories, including anger, sadness, joy, disgust and fear. In our previous work [18], a fast Naive Bayes classifier is trained on Weibo data for emotion classification. The system named MoodLens whose vital part is the emotion classifier, is now available online for temporal and spatial sentiment pattern discovery. We arrange the tweets related to stock market, with one day as the time unit, and employ the system to label them with the emotions. There are five online emotion time series: $X_{anger}, X_{sadness}, X_{joy}, X_{disgust}$ and X_{fear}. Online emotions are represented by $X = (X_{anger}, X_{sadness}, X_{joy}, X_{disgust}, X_{fear})$.

Observing the time series, the volume of tweets reduces significantly on non-trading days. Figure 1 shows the volume of the tweets related to the stock market from September 1st to 16th in 2015. There are separately Memorial Day between September 3rd and 5th and a weekend between September 12th and 13th, which are both non-trading days. We consider that online stock market emotion on non-trading days could not help us analyze and predict Chinese stock market. Hence, removing the data items on non-trading days from the time series, the results retain significant emotion data. It also partly reflects that tweets selected by the keywords could represent the stock market.

For the sake of stability of online emotion data, we measure the relative value (proportion) of each mood on one day as the final online stock market emotion X. Figure 2 shows online stock market emotion time series X from September 1st to 16th in 2015. We observe the spike in X_{anger} on May 12th in 2015, when there is a heated argument between CEOs of listed companies. On June 19th in 2015, there was a plunge in Chinese stock market, with a fall of the index with

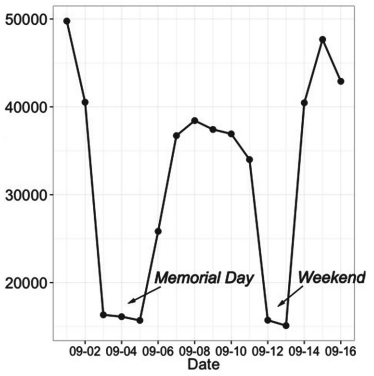


Fig. 1. Volume of the tweets related to the stock market from September 1st to 16th in 2015. There are respectively Memorial Day day between 9–3 and 9–5 and a weekend between 9–12 and 9–13, which are also non-trading days.

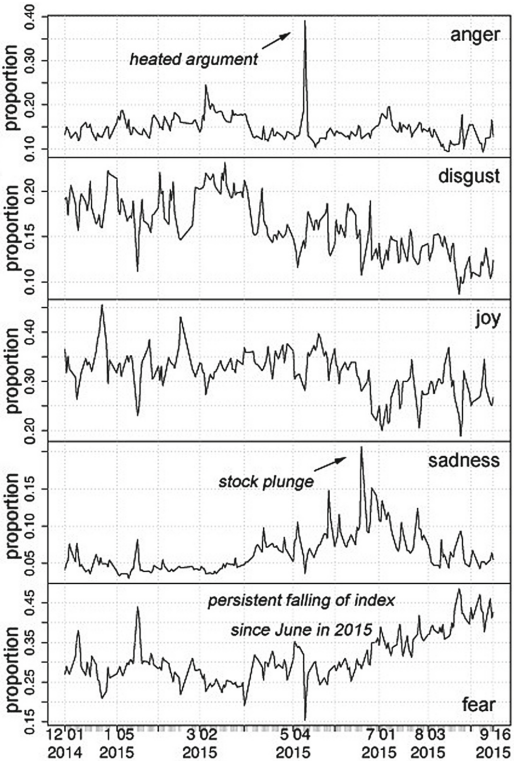


Fig. 2. Time series of each online stock market emotion from December 1st 2014 to September 16th 2015.

6.41 % and $X_{sadness}$ arrived the maximum. Since June in 2015, persistent falling of the index caused inward fears of investors, which can be seen from the sharp growth of X_{fear} in Fig. 2.

From the above observations, it can be concluded that the fluctuation of the sentiments can be connected with remarkable events in the stock market. It further inspires us to investigate the correlation and even causality between emotions and the market, which will provide the foundation for the predicting models.

3.2 Stock Market Data

In China, the economists and traders regard the Shanghai Stock Exchange Composite Index as reflecting the overall status of the Chinese stock market. Therefore, the index is selected as price attribute of the stock market to analyze and predict. In particular, there are four values in candlestick charts of the index, which are respectively the closing index, the opening index, the intra-day highest index, the intra-day lowest index. We transform the values of the index into *Close*, *Open*, *High* and *Low* (to express rate of change on i-th day), and they can be written as

$$\begin{aligned} Close_i &= \frac{Index_{close,i} - Index_{close,i-1}}{Index_{close,i}} \times 100, \\ Open_i &= \frac{Index_{open,i} - Index_{close,i-1}}{Index_{close,i}} \times 100, \\ High_i &= \frac{Index_{high,i} - Index_{close,i-1}}{Index_{close,i}} \times 100, \\ Low_i &= \frac{Index_{low,i} - Index_{close,i-1}}{Index_{close,i}} \times 100. \end{aligned} \tag{1}$$

In addition to these four attributes, the trading volume of Shanghai Stock Exchange is also a key target used to reflect the status of the Chinese stock market. The time series of trading volume on each day is not transformed at all.

We crawl historical data of the index and trading volume from December 1st 2014 to December 7th 2015. In this period, the number of trading days is totally 249 in our research. As a result, we obtain five time series which depict stock market's state on each day including $Y_{close}, Y_{open}, Y_{high}, Y_{low}$ and Y_{volume}. Each time series is a column vector of Y (shown in Fig. 4), i.e., $Y = (Y_{close}, Y_{open}, Y_{high}, Y_{low}, Y_{volume})$.

The dataset (X and Y) is divided into two parts according to the date: the 80 % data for training (from December 1st 2014 to September 16th 2015) and the 20 % data for testing (from September 17th to December 7th in 2015). The training set is used to not only analyze the relation between online emotions and the stock market but also fit and estimate the prediction model. The testing set is kept in a vault and brought out only at the end of evaluation in realistic application.

4 Correlation Between Online Emotions and the Stock Market

The preceding part of the paper describes the two groups of time series (in the training set): X (represents online stock market emotions) and Y (represents the stock market), which contribute to discuss the correlation between online emotions and the stock market. However, the purpose of the paper is to find out whether online emotions can predict the stock market in China. Supposing that online emotions ahead of 1 to 5 days are available for stock prediction, we shifted emotion series to an earlier date: 1 to 5 days. Hence, each emotion corresponds to 5 time series according to shifted time. Each category of online emotions can be defined as (the categories of emotions are represented by e, $e = anger, sadness, joy, disgust$, or $fear$) $X_e = (X_{e,1}, X_{e,2}, X_{e,3}, X_{e,4}, X_{e,5})$.

For the analysis the relation of X and Y (T represents one certain time series of X or Y), we normalize all the time series, of which data items are transformed to the values from 0 to 1 as

$$T_i = \frac{T_i - T_{min}}{T_{max} - T_{min}}, \tag{2}$$

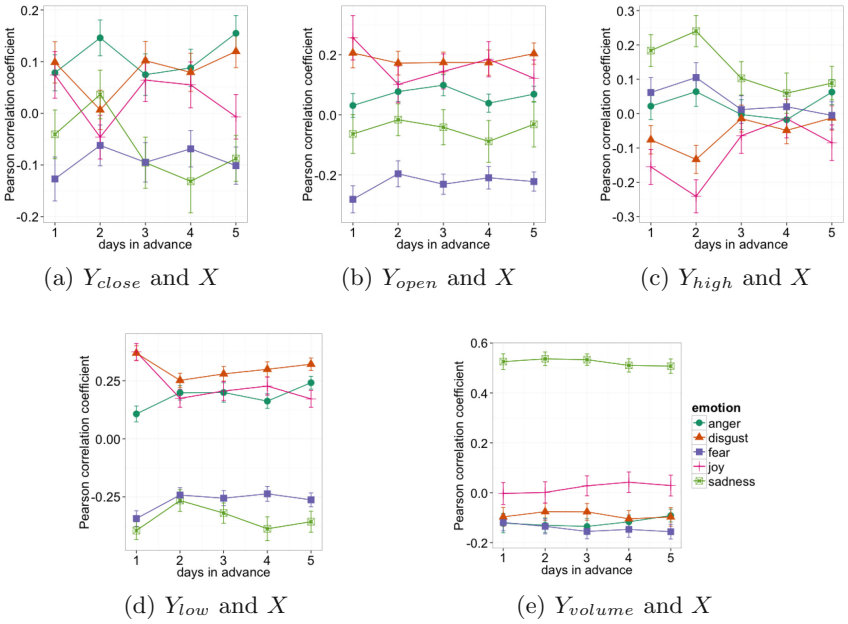
T_i is the i-th item in time series T, T_{max} is the maximal value of T, and T_{min} is the minimal value of T. Then, by using Pearson correlation analysis, we measure the linear dependence between x ($X_{e,t}$, the emotion e ahead of t days in X) and y (one target of Y) as Eq. (3). ρ is the Pearson correlation coefficient of time series x and y defined as

$$\rho = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}. \tag{3}$$

For observing whether there are distinct differences of correlation coefficient between online emotions and the stock market, the emotion time series associated with stock market time series are sampled 100 times. In one time, we sample randomly 150 pairs (from 191 pairs in the training set) of data items respectively from emotion time series and stock market time series. We calculate 100 sampling results' correlation coefficient, and then obtain the mean values and standard deviations. Figure 3 shows the means and error bars, which depicts sampling results' correlation coefficient. It can be seen that there are significant differences between different emotions.

In addition, we randomly shuffle the time series 100 times and calculate the Pearson correlation coefficient of them. Comparing the mean coefficients with that of non-shuffled time series, we find that all the correlation coefficients (shuffled) are near 0, and it suggests that most of correlation coefficients (not shuffled) are relatively higher than random results, indicating the significance of the correlation we find.

Through the correlation coefficients above (shown in Fig. 3), we set the threshold of correlation coefficient ρ as 0.2 (the absolute value) and find some interesting and valuable results. The correlation between all online emotion time



(a) Y_{close} and X (b) Y_{open} and X (c) Y_{high} and X

(d) Y_{low} and X (e) Y_{volume} and X

Fig. 3. Pearson correlation coefficient between five targets of the stock market and online emotion time series.

series (in X) and Y_{close} is very low ($\rho < 0.2$), which indicates little linear dependence between them. As to Y_{open}, the correlation coefficients with X_{fear} (ahead of 1, 3, 4 and 5 days), X_{joy} (ahead of 1 day) and $X_{disgust}$ (ahead of 1 and 5 days) are more than 0.2. Y_{open} is negatively correlated with X_{fear}, positively correlated with X_{joy} and $X_{disgust}$. As to Y_{high}, the correlation coefficients with X_{joy} (ahead of 2 days) and $X_{sadness}$ (ahead of 2 days) are more than the threshold. Y_{high} is negatively correlated with X_{joy}, positively correlated with $X_{sadness}$. Y_{low} and 5 types of emotion time series have relatively high correlation, and the correlation coefficients between Y_{low} and $X_{sadness}$ (ahead of 1 and 4 days) is the highest ($|\rho| > 0.4$). Y_{low} is negatively correlated with X_{anger}, $X_{disgust}$ and X_{joy}, positively correlated with X_{sad} and X_{fear}. An interesting finding is that the correlation between Y_{volume} and $X_{sadness}$ (no matter ahead of how many days) is unexpectedly high, correlation coefficients ρ of which is more than 0.5. Besides, Y_{volume} and other online emotion time series don't have a comparatively strong ($\rho > 0.2$) correlation.

5 Granger Causality Test of Online Emotions and the Stock Market

Despite the correlation analysis, we also preform the causality test further on the training data. Here we apply the econometric approach named Granger causality

test to study the relation between online emotions and the stock market. The Granger Causality Test is a statistical hypothesis test for determining whether one time series is functioning in forecasting another. One time series x is said to Granger-cause y if it can be shown that x provides statistically significant information about future values of y, usually through a series of t-tests and F-tests on lagged values of x. We perform the analysis according to models shown in Eqs. 4 and 5 for the period from December 1st 2014 to September 16th 2015.

$$y_t = \alpha + \sum_{i=1}^{n} \beta_i y_{t-i} + \epsilon_t, \tag{4}$$

$$y_t = \alpha + \sum_{i=1}^{n} \beta_i y_{t-i} + \sum_{i=1}^{n} \gamma_i x_{t-i} + \epsilon_t. \tag{5}$$

The Granger causality test could select only two time series as inputs. We apply Granger causality test respectively on two groups: online emotion time series and stock market time series. Delaying time is set to 1, 2, 3, 4 and 5 days. According to different delaying time, we calculate the p-value to determine the results of hypothesis test. Here, significance level is set to 5 %.

We list testing results whose p-values are required to different significant levels in Table 1. According to the results of Granger causality test, the null hypothesis, $X_{disgust}(lag = 1, 2)$ series do not predict Y_{close}, with a high level of confidence (p-value < 0.01) can be rejected. However, the other emotions do not have causal relations with Y_{close}. Y_{open} and X_{joy} (p-value < 0.001), X_{fear} (p-value < 0.001) and $X_{disgust}$ (p-value < 0.05 or even 0.01) have causal relations. X_{joy}, $X_{sadness}$ and $X_{disgust}$ have causal relations with Y_{high} and Y_{low} (p-value < 0.05 or even 0.01). At last, the results also suggest trading volume in stock market time series do not have significant causal relation with any emotion time series (p-value ≥ 0.05). It's surprising to find that X_{anger} in online emotion time series does not have causal relation with any attribute of stock market in China.

The above analysis shows that $X_{disgust}$, X_{joy}, $X_{sadness}$ and X_{fear} can be promising features for the stock prediction models, except for X_{anger}.

6 Predict the Stock Market

Firstly, in this section, based on discretization methods, regression problems of predicting the stock market are converted to corresponding classification problems. Next, we perform linear and non-linear methods to solve the classification problems of stock market prediction. Eventually, the classification models are validated by 5-fold cross-validation on training set and we obtain a group of high-performance prediction models named SVM-ES.

For the prediction issue, we make use of the online emotion time series set (composed by shifted time series with different lags ranging from 1 to 5 for five emotions) or its subsets within the period from December 1st 2014 to September 9th 2015. Setting the longest lag to 5 trading days, the actual stock market time series are Y from December 8th 2014 to September 16th 2015.

Table 1. Results of Granger causality test of online emotion and stock market time series. Only significant results are listed because of the limited space. p-value < 0.05: *, p-value < 0.01: **, p-value < 0.001.

emotion	lag (days)	Close	Open	High	Low	Volume
anger	1					
	2					
	3					
	4					
	5					
disgust	1	0.0057^{**}			0.0322^{*}	
	2	0.0062^{**}				
	3		0.0067^{**}			
	4		0.0190^{*}			
	5			0.0280^{*}		
joy	1		0.0005^{***}	0.0234^{*}		
	2		$6.e-5^{***}$	0.0304^{*}		
	3		$2.e-5^{***}$	0.0087^{**}	0.0058^{**}	
	4		$7.e-5^{***}$	0.0385^{*}	0.0352^{*}	
	5		0.0006^{***}			
sadness	1			0.0115^{*}	0.0272^{*}	
	2			0.0224^{*}		
	3			0.0303^{*}		
	4					
	5					
fear	1		0.0001^{***}			
	2		$2.e-5^{***}$			
	3		$3.e-6^{***}$			
	4		$6.e-6^{***}$			
	5		0.0002^{***}			

6.1 Discretization of Stock Market Data

As illustrated in the previous sections, $Y_{close}, Y_{open}, Y_{high}$, Y_{low} and Y_{volume} are our targets of prediction in the stock market. Investors always just care for whether $Y_{close,i}$ (the element on i-th day in Y_{close}) are positive or negative, which will help investors make decisions to conduct stock transactions, and the binary classification (positive or negative) of Y_{close} and Y_{open} are also the part of our targets for prediction.

Besides, we convert regression problems of predicting five attributes in the stock market to classification problems by discretization methods through which we classify each of attributes to three categories. Specifically, $Y_{close}, Y_{open}, Y_{high}$, and Y_{low} are divided into three categories: **bearish**(-1), **stable**(0) and **bullish**(1) represented by CLOSE, OPEN, HIGH and LOW below. Y_{volume} are divide into three categories: **low**(-1), **normal**(0) and **high**(1) represented by VOLUME below.

The discretization of five attributes in the stock market is conducted by two methods: equal frequency and K-means clustering. Equal frequency discretization is a simple but effective method that we sort items from large to small then cut them into 3 clusters of even size. K-means clustering, another method we use, is popular for cluster analysis in data mining. In this paper, K-means clustering aims to partition observations of the stock market into 3 clusters in which each observation belongs to the cluster with the nearest distance. The results of three categories discretization by K-means are shown in Fig. 4 with 3 different grey levels.

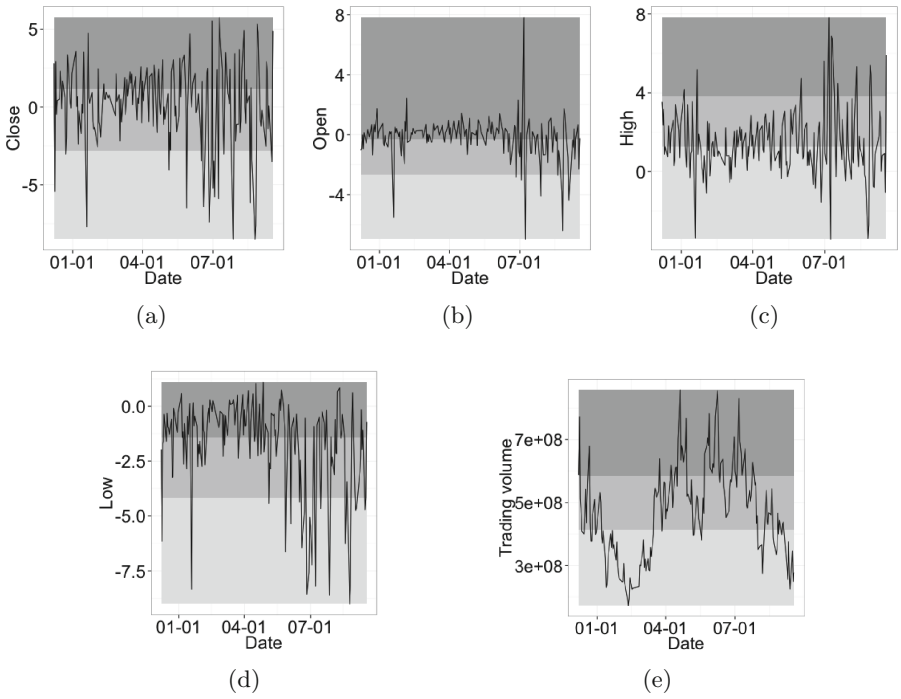


Fig. 4. Stock market time series and discretization results (by K-means) of $Y_{close}, Y_{open}, Y_{high}$, Y_{low} and Y_{volume}.

Intuitively, as compared to the approach of equal frequency, the discretization based on K-means is more flexible and adjustable to the dynamic of the market. The categories it generates can better reflect the actual market status and thus can offer us a better benchmark to test the prediction results.

6.2 Classification Model for Stock Prediction

In this paper, we perform machine learning methods, Logistic Regression (linear) and Support Vector Machine (non-linear), to solve the classification problems for

stock prediction. These methods are both popular for training binary or multiple classification. To predict the categories $(-1, 0, 1)$ or $(0, 1)$ (just for CLOSE and OPEN) of Y on i-th day, the input attributes of our Logistic Regression model (LR) and Support Vector Machine model (SVM) include only online emotion values of the past 5 days or a subset of them, except for other variables in the field of finance. We adapt 5-fold cross-validation to examine the accuracies of models.

At the outset, we consider all five emotions of the past 5 days as the input attributes of LR and SVM. The accuracies of models by 5-fold cross-validation are shown in Table 2 (3-categories and 2-categories). For the classification problem in this paper, the performance of SVM is always better than that of LR. Therefore, we conjecture that, relation between online emotions and the stock market is not simply linear, and the relation is more likely complicated and nonlinear.

While 3-categories discretization $(-1, 0, 1)$ results in the stock market as predicted targets, K-means clustering is always better than equal frequency discretization. In other words, the accuracies of models by 5-fold cross-validation, of which predicted targets are the results by K-means clustering, are relatively higher. Considering the categories generated by K-means discretization better represent the market status, we can conclude that our models indeed capture the essence of the stock fluctuation.

However, recalling the correlation analysis and Granger causality test of online emotions and the stock market, not all the emotions play roles on predicting the stock market and the analysis results should be used for the feature selection. Consequently, we build support vector machine model based emotions selected (SVM-ES) for stock prediction (discretized by K-means). The input attributes are based on analysis results of Granger causality test and Pearson correlation. We select $X_{disgust}$ (ahead of 1, 2 days) for the SVM-ES to predict CLOSE, X_{fear} (ahead of

Table 2. Accuracies of 5-fold cross-validation for 3-categories and 2-categories prediction models.

Target (3)	equal frequency		K-means		
	LR	SVM	LR	**SVM**	**SVM-ES**
CLOSE	34.0%	43.5%	52.9%	**58.1**%	57.6%
OPEN	37.7%	44.0%	53.4%	61.3%	**64.4**%
HIGH	36.7%	39.3%	48.7%	53.4%	**54.5**%
LOW	42.4%	49.2%	57.0%	63.4%	**64.4**%
VOLUME	50.8%	63.9%	53.4%	**67.0**%	66.5%

Target (2)	LR	SVM	SVM-ES
CLOSE	58.1%	**61.3**%	60.2%
OPEN	58.1%	**66.0**%	64.9%

1–5 days), X_{joy} (ahead of 1–5 day) and $X_{disgust}$ (ahead of 3 and 4 days) as the input attributes for predicting OPEN, X_{joy} (ahead of 1-4 days), $X_{sadness}$ (ahead of 1–3 day) and $X_{disgust}$ (ahead of 5 days) as the input attributes for predicting HIGH, and $X_{sadness}$ (ahead of 1 day), X_{joy} (ahead of 1–3 day) and $X_{disgust}$ (ahead of 5 days) as the input attributes for predicting LOW. Correlation analysis of Y_{volume} indicates that Y_{volume} and $X_{sadness}$ (ahead of 1–5 days) have the strongest correlation ($\rho > 0.5$) among all online emotions, however, just using sadness as the learning feature surprisingly can not guarantee the expected performance. Thus, we try to select $X_{sadness}$ (ahead 1–5 days) and X_{fear} (ahead 1–5 days) which is the second strongest relation with Y_{volume} as the input attributes to predict VOLUME.

After adjusting and fixing the input attributes of SVM-ES, we train the models for stock prediction. The last column of Table 2 shows the accuracy of 5-fold cross-validation, respectively for 3-categories and 2-categories classification models. There are slight differences in performance between SVM-ES and the SVM trained using all the emotions, indicating emotions selected are playing dominant roles in forecasting the market. It is noteworthy that input attributes of all the SVM-ES don't include anger and it's surprising that anger shown in online social media possesses the weakest correlation or even no relation with the Chinese stock market.

From Table 2 it should be also noted that, emotions selected can boost the classification results attributes like OPEN, HIGH and LOW, while for CLOSE and VOLUME, SVM with all emotions as features is still the most competent solution, with slight increment (around 1 %) to SVM-ES (few attributes of input). This result explains that emotions except for input attributes of SVM-ES have very weak effects on the stock market prediction.

6.3 Evaluation in Realistic Application

For further evaluating our prediction models, we sustain collecting stock-relevant tweets on Weibo with APIs and process them so as to obtain online emotion time series as our testing set from September 17th to December 7th in 2015. Then we apply our classification models SVM-ES for stock prediction in the realistic Chinese stock market and we can get the daily predictions of five attributes before the market open. Framework of realistic application based on SVM-ES is demonstrated in Fig. 5. We evaluate the stock market prediction application and the accuracies are shown in Table 3. It turns out that the model achieves the high prediction performance, especially with accuracy of 64.15 % for the intraday highest index (3-categories) and the accuracy of 60.38 % for the trading volume (3-categories).

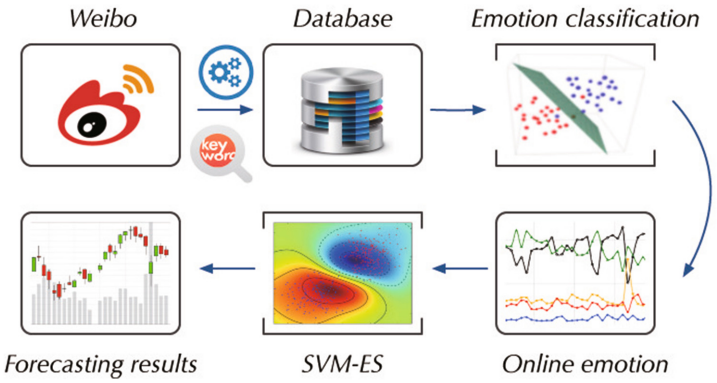


Fig. 5. Framework of realistic application for stock prediction based on SVM-ES.

Table 3. Accuracies of SVM-ES on realistic application.

CLOSE (3)	OPEN (3)	HIGH (3)	LOW (3)	VOLUME (3)	CLOSE (2)	OPEN (2)
56.60 %	43.40 %	64.15 %	56.60 %	60.38 %	60.38 %	56.60 %

7 Conclusion

In this paper, we collect massive tweets in Weibo with five categories of sentiments and focus on the stock market in China. The correlation analysis and Granger causality test are performed, which suggest that several emotions can be directly used to predict the market. Based on this, we establish several models to predict the closing index, the opening index, the intra-day highest index, the intra-day lowest index and trading volume. The results show that our model SVM-ES can outperform baseline solutions. Finally, we also testify its performance in the realistic application. In conclusion, our findings in this paper confirm that the stock market in China can be predicted by various online emotions including disgust, joy, sadness and fear.

This study has inevitable limitations, which might be interesting directions in the future work. For example, the detailed connection between the emotion and the market still remains unclear and how it evolves with time is also not discussed, however, which could help to design incremental learning schemes.

References

1. Baker, M., Wurgler, J.: Investor sentiment in the stock market. Working Paper 13189, National Bureau of Economic Research, June 2007
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)

3. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A.: Web search queries can predict stock market volumes. PLoS ONE **7**(7), e40014 (2012)
4. Brown, G.W., Cliff, M.T.: Investor sentiment and the near-term stock market. J. Empir. Financ. **11**(1), 1–27 (2004)
5. Cohen-Charash, Y., Scherbaum, C.A., Kammeyer-Mueller, J.D., Staw, B.M.: Mood and the market: can press reports of investors' mood predict stock prices? PLoS ONE **8**(8), e72031 (2013)
6. Dolan, R.J.: Emotion, cognition, and behavior. Science **298**(5596), 1191–1194 (2002)
7. Gayo-Avello, D.: "I wanted to predict elections with twitter and all i got was this lousy paper"-a balanced survey on election prediction using twitter data. arXiv preprint arXiv:1204.6441 (2012)
8. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical report, Stanford Digital Library Technologies Project (2011)
9. Mao, H., Counts, S., Bollen, J.: Quantifying the effects of online bullishness on international financial markets. In: ECB Workshop on Using Big Data for Forecasting and Statistics, Frankfurt (2014)
10. Nofsinger, J.R.: Social mood and financial economics. J. Behav. Financ. **6**(3), 144–160 (2005)
11. Oh, C., Sheng, O.: Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In: Galletta, D.F., Liang, T.P. (eds.) ICIS. Association for Information Systems (2011). http://dblp.uni-trier.de/db/conf/icis/icis2011.html#OhS11
12. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC, vol. 10, pp. 1320–1326 (2010)
13. Parikh, R., Movassate, M.: Sentiment analysis of user-generated twitter updates using various classification techniques. Technical report (2009)
14. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
15. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Trans. Inf. Syst. **27**(2), 1–19 (2009)
16. Wanyun, C., Jie, L.: Investors' bullish sentiment of social media and stock market indices. J. Manag. **5**, 012 (2013)
17. Ding X., Zhang Y., Liu T., Duan J.: Deep learning for event-driven stock prediction. In: IJCAI, pp. 1–7, July 2015
18. Zhao, J., Dong, L., Wu, J., Xu, K.: Moodlens: an emoticon-based sentiment analysis system for Chinese tweets. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1528–1531. ACM (2012)