Professor Yong SHI, PhD
E-mail:yshi@ucas.ac.cn
Ye-ran TANG, PhD Candidate
E-mail:tangyeran14@mails.ucas.ac.cn
Ling-xiao CUI, PhD
E-mail: clxyx@itp.ac.cn
Associate Professor Wen LONG*, PhD
E-mail: longwen@ucas.ac.cn(*Corresponding author)
School of Economics & Management, University of Chinese Academy of Sciences
Research Center on Fictitious Economy & Data Science Chinese Academy of Sciences
Key Laboratory of Big Data Mining & Knowledge Management Chinese Academy of Sciences

A TEXT MINING BASED STUDY OF INVESTOR SENTIMENT AND ITS INFLUENCE ON STOCK RETURNS

Abstract*. This paper studies a broad sample of investors' online opinion posts on the largest stock forum in China. Using text mining methods with data cleaning, text representation, feature extraction, and two-step sentiment classification, the paper identifies individual investor sentiment and complies an index. Further, the investor sentiment index is applied to Chinese stock market, and a relatively comprehensive analysis is proposed to study the relationship between investor sentiment and the CSI 300 stock index returns. Empirical results suggest prediction effect of investor sentiment on the stock market returns. Investor sentiment has short-term positive effects and medium-run reverse effects on stock market. The asymmetric effect that high investor sentiment gets more obvious influence on stock returns has also been found. We examine cumulative changes of investor sentiment to verify our main conclusion.*

Keywords*: investor sentiment, stock returns, financial market, text mining, stock forum*

JEL Classification: G10, C65

DOI: 10.24818/18423264/52.1.18.11

1. Introduction

The influence of investor sentiment on the stock market has presented puzzles. The stock pricing model based on hypothesis of rational economic man suggests that the price of a stock is determined by the discounted future dividends, and investor sentiment will have little influence on the stock market. However, researches on behavioral finance show that there are several irrational phenomena stock market, and this irrational deviation is systemic (Black, 1986; Long et al., 1991). Investor sentiment is one of the important factors in these irrational factors. The vast majority of the sentiment literature suggests individual investors are sentiment traders (Lee et al., 1991; Barberis and Xiong, 2010), and these sentiment traders have impact on stock price.

To study investor sentiment and its influence on stock returns, the data that can reflect investors' opinions on future trend are needed. In previous literatures, investor sentiment indicators mainly can be divided into three categories: the investor sentiment index extracted from surveys of consumer and investor confidence, indirect indicators obtained by the historical trading data of the stock market, and investor sentiment index obtained from the network information.

The studies with surveys mainly use a variety of consumer sentiment survey index and consumer confidence index as a proxy variable of investor sentiment(Otoo, 1999; Lemmon and Portniauina, 2006). Otoo (1999) use the University of Michigan consumer survey sentiment index as investor sentiment and find stock prices can influence consumer sentiment but the opposite is not significant. Lemmon and Portniaguina (2006) use two surveys of consumer confidence conducted by the Conference Board and the University of Michigan Survey Research Center, and find that investor sentiment can predict stock returns in the short term, but cannot predict long-term changes in stock market. The conclusions of these studies suggest investor sentiment has impact on stock market, especially in short term. However, investor sentiment index extracted from surveys takes a lot of time and cost for survey so that it is hard to get higher frequency data to test such short-term effects.

The indirect indicator of investor sentiment mainly obtained from the historical trading data (Wheatley and Neal, 1998; Baker et al., 2012; Stambaugh et al., 2012, 2015; Berger and Turtle, 2015; Chau et al., 2016). Bakeret al adopt a composite index extracted from six variables of the stock market with principal component analysis method to represent investor sentiment. They find that high

DOI: 10.24818/18423264/52.1.18.11

investor sentiment is often accompanied with large arbitrage risk and low subsequent stock returns, while low investor sentiment has reversed effect. Compared with ways of surveys, using historical trading data ofstock market as proxy variable indeed saves time and effort, but it is an indirect indicator of sentiment and cannot reflect investors' opinion directly. Besides, because the investor sentiment is extracted from historical data of stock market, it can hardly reflect the effects of new information.

Information on the Internet contains abundant financial market content, and can also reflect investor sentiment orientation. Thanks to the high penetration of the Internet, investors can freely post their opinions about the stock market on the Internet, and stock forum has collected a large number of investors' expectations. Besides, in recent years, the development of data mining technology provides convenience for processing and analyzing massive text data. Since these information on stock forum can directly represent investor sentiment and be easily collected, applying text mining technology to obtain individual sentiment on stock forum to analyze its influence on stock markets has attracted much attention (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Das and Chen, 2007; Sun et al, 2016; Arvanitis and Bassiliades, 2017).Arvanitis and Bassiliades (2017)use Naïve Bayes classifier and the n-gram probabilistic language model to create a sentiment index and find a twice predictive ability of Baker and Wurgler's index. Compared with indicators obtained from the historical trading data, investor sentiment on stock forum can directly reflect individuals' opinions on the future trend of the stock so that it will have much more influence on stock price. Compared with consumer confidence index, using data on stock forum as investor sentiment, we can get higher frequency data to test its short-term effects. However, the previous literatures based on stock forum sentiment mainly have following problems: (1) A large number of noisy posts are involved in stock forum, which makes it difficult to improve accuracy of sentiment classification; (2) Only reflect the phenomenon of the relationship between investor sentiment and stock returns, but lack of mechanism research.

In this paper, with the large-scale online stock forum data, we employ a two-step text mining methods to identify individual investor sentiment, which to eliminate noise and to improve the accuracy of distinguishing investor sentiment in order to better test of the relationship between investor sentiment and stock market returns. We apply our sentiment recognition method to the most popular

DOI: 10.24818/18423264/52.1.18.11

popular stock forum in China, and compose the investor sentiment index to verify effect on CSI 300 index. We find investor sentiment has a short-term positive and medium-run reverse effect on stock market due to the stock mispricing (overvalue or undervalue) caused by sentiment, which is different from previous studies whose main conclusion is a negative relationship between investor and stock returns. Further, recent findings suggest that consumer sentiment-based indicators have less value in describing equity market dynamics (Berger and 2015) because institutions are the primary sentiment traders. However, our finding argues a significant effect of individual sentiment on stock returns.

2. Data and sentiment measures

2.1. Investor sentiment measures

We extract individual investor sentiment from large-scale online stocks forum posts on East money stock forum, which is one of the largest stock forum in China. East money is a financial portal founded in 2004, and its stock forum's and posts have reached a certain scale since the year of 2011. It has become the largest and the most influential financial portal whose effective visit time is accounted for 43.8% of total effective visit time in financial portals according to several authoritative surveys. We use 5,163,210 online posts on East money stock forum from November 1st, 2011 to September 30th, 2015 for sentiment measurement.

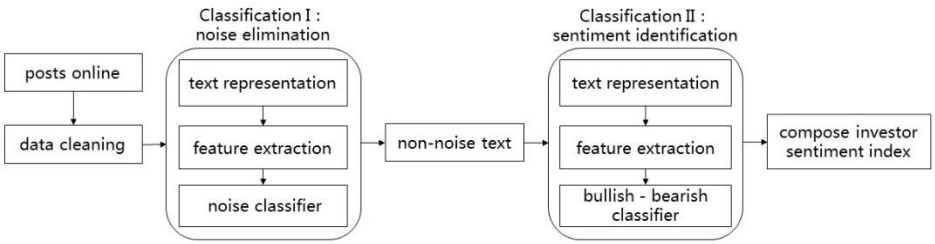


Figure 1. Text mining process for investor sentiment measures

With such a large scale of the posts data, we apply sentiment analysis technology to automatically classify unstructured reviews as positive or negative, and then identify investor sentiment as either bullish or bearish. Considering online posts involves a large number of noise which has their own characteristics, we use two-step classifier for sentiment analysis. The first step is to eliminate noise by dividing the text into noise and non-noise text, and the second step is to divide

DOI: 10.24818/18423264/52.1.18.11

non-noise text into bullish and bearish for sentiment identification. This two steps will use several technologies including data cleaning, text representation, feature extraction and classification. After that, we combine individual sentiments into investor sentiment index and apply it into stock market. Text mining process for investor sentiment measures is shown in Figure 1.

(1) Data cleaning

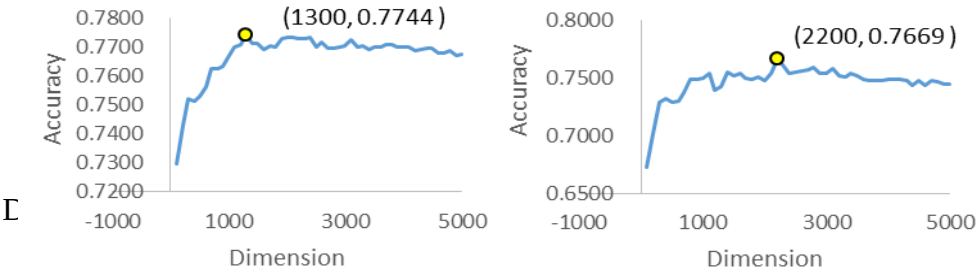
Online stock forum gathers a large number of investors' posts. The text data there contain a lot of punctuation, noise, etc., and cannot directly use for analysis. Therefore, we use data cleaning technology to eliminate punctuations and gibberish, and to preprocess text such as word segment.

(2) Text representation

Text representation is a technology to change text information into digital data for computer processing. Vector space model (Salton et al., 1975) is one of the text representation methods based on and extending the bag-of words model, and it is commonly used in methods of text classification. The bag-of-words model represents a text as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. Since forum data have a lot of new words and nonstandard sentences, it is suitable for the hypothesis that each word is independent in vector space model. With vector space model, the texts are expressed as a vector or a point in the language space, and each word gets the weight according to its importance in the text.

(3) Feature extraction

If the total number of the words in sample is large and each text only involves a part of the words, feature extraction method will be used to remove the words that contribute little for analysis, which may reduce computing complexity and avoid over fitting. According to previous studies (Blum and Langley, 1997), we use CHI statistics for feature extraction, and decide feature dimension by data experiment, which is to calculate the classification accuracy in different dimensions respectively. The results are shown in Figure 2. It suggests that classification accuracy increases with increase of the feature dimension, however, after it reaches a certain level, increasing feature dimension can no longer improve the classification accuracy. As a result, we choose 1300 dimensions of feature for the first step classification to eliminate noise text, and choose 2200 dimensions of feature for the second step classification for sentiment identification.



(I) Noise elimination (II) Sentiment identification

Figure 2. Classification accuracy in different dimensions

(4) Classifiers for sentiment identification

We use Support Vector Machine (SVM) to identify sentiment, which has already become the main approach for classification recently due to its performance (Cortes and Vapnik, 1995; Deng et al., 2012). We decide size of training set by data experiment, which is to calculate the classification accuracy in different size of training set respectively. The results are shown in Table 1. It can be seen that the accuracy of first-step classification becomes stable when the size of the training set is more than 4500, and the accuracy of second-step classification becomes stable when the size of the training set is more than 1600. Considering both accuracy and convenience, we use training set whose size is 6091, including 1216 bullish text, 1011 bearish text, and 3864 noise text. We use 10-fold cross-validation method to calculate our accuracy, and get the accuracy of 77.44% for the first step classification to eliminate noise text, and get the accuracy of 76.69% for the second step classification for sentiment identification.

Table 1. Classification accuracy in different training set size

Classification I：Noise elimination						
Training set size	3000	4000	4500	5000	5300	5500
Accuracy	76.07%	76.15%	77.16%	77.13%	77.79%	77.29%
Classification II：Sentiment identification						
Training set size	1000	1200	1400	1600	1800	2000
Accuracy	73.73%	74.76%	75.39%	76.69%	76.74%	76.48%

(5) Compose investor sentiment index

In order to combine the identified investor sentiment into investor sentiment index, we reference the method in previous literatures (Antweiler and Frank, 2004), define M_t^{BUY} as total bullish posts in time interval *t*, and M_t^{SELL} as total bearish posts in time interval *t*. The formula for sentiment index composition is:

$$M_t = \ln\left[\frac{1+M_t^{BUY}}{1+M_t^{SELL}}\right] \quad (1)$$

Based on 5,163,210 online posts on East money stock forum, we use text mining methods mentioned above to compose investor sentiment index. Figure 3

DOI: 10.24818/18423264/52.1.18.11

shows the daily investor sentiment index in a recent year.

It can be seen that investor sentiment is in the interval of [-0.8, 0.8], and has relatively obvious trends in different periods. The mean of investor sentiment index from November 1st, 2011 to September 30th, 2015 is 0.054, which is more than 0 slightly. It suggests that on average, investor sentiment is towards bullish, which verifies the viewpoint of investors' irrational biases in behavioral finance (Odean, 1998). Previous literatures related to behavioral finance suggest that due to the cognitive bias of overconfidence, investors often overestimate their chances of success. Stambaugh et al. (2012) have also explained this phenomenon from another aspect. They believe that high investor sentiment means the overvalued of the stock price, and as short-selling restrictions hamper the overvalued stock price to return to its value, investor sentiment is more easily to keep high level.

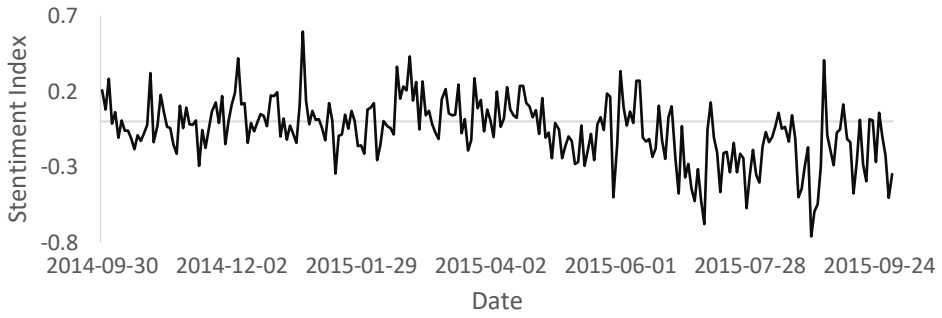


Figure 3. Investor sentiment index from Sep. 30th, 2014 to Sep. 30th,2015

2.2. The relationship between investor sentiment and stock market returns

We use returns of CSI 300 index to reflect the stock market returns in China. Define the price of the stock index at time *t* as p_t, and then the return of the stock index at time *t* is $R_t = \ln p_t - \ln p_{t-1}$. We use both weekly and daily data of CSI 300 index from November 1st, 2011 to September 30th, 2015.

We analyze the correlation of bullish or bearish between investor sentiment index and CSI 300 index. Here we further introduce the concept of institutional view for comparison, which is denoted as S_t. We use the rating data of institutions from November 1st, 2011 to September 30th, 2015 in WIND database to indicate institutional view on stock market. Define total bullish institutions in time interval *t* as S_t^{BUY}, and define total bearish institutions as S_t^{SELL}. The formula for institutional view S_t is:

DOI: 10.24818/18423264/52.1.18.11

$$S_t = \ln\left[\frac{1+S_t^{BUY}}{1+S_t^{SELL}}\right] \tag{2}$$

For individual investor sentiment, define bullish sentiment when $M_t > 0$, bearish sentiment when $M_t < 0$, and neutral sentiment when $M_t = 0$. For institutional view, define bullish view when $S_t > 0$, bearish view when $S_t < 0$, and neutral view when $S_t = 0$. For stock market, define bullish market when $R_t > 0$, bearish market when $R_t < 0$, and sideways market when $R_t = 0$ .The same sign of M_t and R_t suggests the similarity between individual investor sentiment and stock market, and the same sign of S_t and R_t suggests the similarity between institutional view and stock market.

We compare the accuracy of bullish or bearish identification between individual investor sentiment and institutional view in 953 trading days from November 1st, 2011 to September 30th, 2015. The similarity rate of institutional view is 50.37%, and t-value is 0.227. The similarity rate of investor sentiment is 60.76%, and t-value is 6.796, which suggests the accuracy of investor sentiment is more than 50% significantly, while the result of institutional view is no more than flipping a coin. A reasonable explanation is that institutional view prefer to quarterly or annual prediction of the stock price, while our individual investor sentiment focus on capturing the sentiment changes in real time. This result suggests that investor sentiment from online stock forum can predict stock returns, especially short term.

3. Empirical test

3.1. The prediction effects of investor sentiment

In order to test the influence of online investor sentiment on stock returns, we establish the following 3 order VAR model:

$$\begin{aligned} R_t &= \alpha_1 + \sum_{j=1}^{L} \beta_{1,j} \cdot R_{t-j} + \sum_{j=1}^{L} \gamma_{1,j} \cdot M_{t-j} + \varepsilon_{1t} \\ M_t &= \alpha_2 + \sum_{j=1}^{L} \beta_{2,j} \cdot R_{t-j} + \sum_{j=1}^{L} \gamma_{2,j} \cdot M_{t-j} + \varepsilon_{2t} \end{aligned} \tag{3}$$

R_t is the return of the stock market, which represented by CSI 300 index. M_t is investor sentiment index. The lag order *L* is 3.

We use both weekly and daily data from November 1st, 2011 to September 30th, 2015 for analysis. The results are shown in Table 2.

Table 2. Prediction effects of investor sentiment on stock returns

DOI: 10.24818/18423264/52.1.18.11

Daily	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}	α
R_t	**0.005***	**-0.006****	**0.001**	0.051	-0.081**	-0.006	<0.001
	(1.75)	(-1.98)	(0.45)	(1.38)	(-2.26)	(-0.16)	(0.63)
M_t	0.345***	0.130***	0.050	0.695	-0.566	-0.245	0.026***
	(9.49)	(3.45)	(1.39)	(1.42)	(-1.17)	(-0.51)	(3.38)
Weekly	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}	α
R_t	-0.008	-0.010	0.013	0.140*	0.021	0.100	0.001
	(-0.47)	(-0.58)	(0.82)	(1.82)	(0.28)	(1.319)	(0.43)
M_t	0.419***	0.021	0.146**	0.696*	-0.377	-0.001	0.029**
	(5.48)	(0.26)	(2.01)	(1.95)	(-1.05)	(<-0.01)	(2.26)

The table reports t-value in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.

The results for daily data suggest a prediction effect of investor sentiment when its lag order is 1 or 2, and when lag order is 3, there is no significant prediction effect. The daily result of R_t in first column is positive and significant at 10% level, which indicates a positive effect of investor sentiment on stock returns. If the investor sentiment in a previous period is bullish, then the stock returns will increase, and vice versa. The daily result of R_t in second column is negative and significant at 5% level, which indicates a reverse effect of investor sentiment on stock returns. If the investor sentiment in two days before is bullish, then the stock returns will decrease. The result in third column is insignificant, which suggests the investor sentiment three days before has few effect on stock returns. The results suggest that investor sentiment from online stock forum can predict stock returns in Chinese A-share stock market. In the short term, because of the noise traders, investor sentiment has a positive effect on stock returns, and a high sentiment leads to increasing of the stock returns. In the medium run, because the stock price return to normal, investor sentiment has a reverse effect on stock returns, the higher the investor sentiment, the returns of the stock in the future will be lower, and vice versa. Finally, the impact of current investor sentiment on stock market returns will gradually disappear.

Almost all of the results for weekly data are insignificant, which suggests investor sentiment cannot affect stock returns for a very long time. With the fact that stock price returns to stock value which depends on its discounted future dividends, the effects of investor sentiment will disappear. It can also be seen from

DOI: 10.24818/18423264/52.1.18.11

the equation of M_t that stock returns have little influence on invest sentiment conversely.

3.2. The asymmetric effects of investor sentiment

The empirical results above not only verify that investor sentiment has impact on stock market returns, but also reflect a possible mechanism by which investor sentiment affects stock market returns, that is, high investor sentiment makes stock prices overvalued, and low investor sentiment makes stock price undervalued. To verify this mechanism, we study the asymmetric relationship between investor sentiment and stock returns. According to previous studies, Miller (1977) argues that short-selling restrictions in the stock market hinder the use of the overvalued of the stock price for investors. After that, several researchers study on short-selling restrictions (Scheinkman and Xiong, 2003). Their main opinion is when stock prices are overvalued, it is hard for prices to return to their value due to the short-selling restriction, while when the stock price is undervalued, the rational arbitrageurs will soon make stock prices return to normal. Therefore, in general, stock prices are more likely to be overvalued. According to Stambaugh et al. (2012), because of the limitation of short-sell, the value-return of the overvalued stock is accordingly limited. If investor sentiment affects stock returns by overvaluing or undervaluing stock price, the influence of investor sentiment on stock returns should be asymmetric. High investor sentiment will get more obvious impact on stock returns than low investor sentiment.

To verify the asymmetric effects of investor sentiment on stock returns, we divide sentiment into 2 stages, and study the relationship between investor sentiment and stock returns in high sentiment and low sentiment respectively. The method for stage division is similar as before. Define bullish sentiment when $M_t > 0$, and bearish sentiment when $M_t < 0$.We set up two models for analysis. The one is model (4), which reflects the effect of single variable regression between investor sentiment and stock returns. The other is model (5), which adds investor sentiment index into Fama-French three-factor model (Fama and French, 1993), and analyzes the relationship between investor sentiment and excess returns adjusted by Fama-French three-factor model.

$$R_t = a + b \cdot M_t + \varepsilon_t \tag{4}$$

DOI: 10.24818/18423264/52.1.18.11

$$R_t = a + b \cdot M_t + c \cdot MKT_t + d \cdot SMB_t + e \cdot HML_t + \varepsilon_t \quad (5)$$

R_t is return of the stock market, which represented by CSI 300 index. M_t is investor sentiment index. Control variables are the factors of market, size, and value constructed by Fama and French (1993). MKT_t is the excess return on the stock market. SMB_t is a return spread between small and large firms. HML_t is a return spread between stocks with high and low book-to-market ratios.

We use daily data from November 1st, 2011 to September 30th, 2015 for analysis. The results are shown in Table 3.

Table 3. The effects of investor sentiment in different sentiment stage

		M_t	MKT_t	SMB_t	HML_t	a
All	Model (4)	**0.025*** **	-	-	-	-0.001**
		(12.11)	-	-	-	(-1.77)
	Model (5)	**0.001***	0.984***	-0.412***	0.113***	<0.001
		(1.94)	(151.44)	(-25.92)	(7.26)	(0.27)
Bullish	Model (4)	**0.013*** **	-	-	-	0.001*
		(4.17)	-	-	-	(1.65)
	Model (5)	**0.002*** **	1.018***	-0.397***	0.109***	<-0.001**
		(2.63)	(105.65)	(-17.40)	(5.49)	(-1.99)
Bearish	Model (4)	0.054***	-	-	-	0.004***
		(8.49)	-	-	-	(2.94)
	Model (5)	**<-0.001**	0.967***	-0.403***	0.126***	<-0.001
		(-0.30)	(103.99)	(-17.43)	(5.18)	(-0.79)

The table reports t-value in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.

In table 3, the results of first column reflect the effects of investor sentiment in different sentiment stage. The results in first two rows are significant at 1% and 10% level respectively, which suggests in general, there is a significant relationship between investor sentiment and stock returns. The results in both third and fourth rows are significant at 1% level, and results in sixth row is insignificant, which indicates a significant relationship between investor sentiment and stock returns in bullish stage, and few impact of investor sentiment in bearish stage. It shows that the relationship between investor sentiment and the stock market are mainly from bullish stage, which suggests that high investor sentiment will get more obvious

DOI: 10.24818/18423264/52.1.18.11

impact on stock returns than low investor sentiment. Further, the results verify the mechanism of investor sentiment's influence on stock market returns. Investor sentiment affect stock market returns by making stock price deviate from its value. High investor sentiment makes stock prices overvalued, and low investor sentiment makes stock price undervalued. Due to the limitation of short selling, in the short term, high investor sentiment is more likely to lead to stock price overvalued, and the stock returns increase. However, with the value-return of the stock price in the medium run, the returns of the stock market decrease, until the impact of the investor sentiment on stock returns disappear.

4. Robustness analyses

4.1. Robustness in different time period

For robustness analyses, we use VAR model in model (3) in different time period. We divide the whole time series into three periods equally. The first time period is from November 1st, 2011 to April 11th, 2013, the second time period is from April 12th, 2013 to July 14th, 2014, and the last time period is from July 13th, 2014 to September 21st, 2015. The results are shown in Table 4.

Table 4. Prediction effects in different time period

2012-2013	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	**0.0071****	-0.0024	-0.0019	-0.1279**	-0.0532	0.0808
	(0.02)	(0.44)	(0.52)	(0.04)	(0.37)	(0.17)
M_t	0.2711***	0.1750***	-0.0484	0.1718	-0.1451	-0.1838
	(0.00)	(0.00)	(0.41)	(0.89)	(0.90)	(0.41)
2013-2014	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	-0.0023	**-0.0062***	0.00039	0.0486	-0.0076	-0.0059
	(0.49)	(0.09)	(0.91)	(0.40)	(0.90)	(0.92)
M_t	0.4237***	-0.0735	0.0556	0.3035	0.4646	-0.4961
	(0.00)	(0.25)	(0.34)	(0.77)	(0.65)	(0.61)
2014-2015	M_{t-1}	M_{t-2}	M_{t-3}	R_{t-1}	R_{t-2}	R_{t-3}
R_t	**0.0145****	**-0.0144***	0.0084	0.0841	-0.1101*	-0.0302
	(0.04)	(0.06)	(0.23)	(0.18)	(0.07)	(0.62)
M_t	0.3885***	0.0345	0.1927***	1.0537*	-0.4039	-0.2901
	(0.00)	(0.59)	(0.00)	(0.05)	(0.44)	(0.58)

The table reports t-value in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.

DOI: 10.24818/18423264/52.1.18.11

A significant effect of the investor sentiment can be seen in all of the three time periods, which verifies our conclusion above that there is a short-term positive prediction effect and a medium-run negative effect of investor sentiment on the stock returns. In particular, in the last period, both short-term positive effects and medium-run reverse effects are significant.

4.2. Further evidence

We also examine cumulative changes of investor sentiment and analyze the robustness of our above conclusions about the relationship between investor sentiment and stock returns. The method to calculate cumulative changes of sentiment is referenced to Berger and Turtle (2015). We define Sum_{t,M_t+} as the sum of successive increased sentiment during time period *t*.

$$Sum_{t,\Delta M_t+} = \begin{cases} Sum_{t-1,\Delta M_t+} + \Delta M_t, & \Delta M_t > 0 \\ 0, & \Delta M_t \le 0 \end{cases} \quad (6)$$

Where $\Delta M_t = M_t - M_{t-1}$. For the initial time *t=0* in the sample, we set $Sum_{0,\Delta M_t+} = 0$.

According to Berger and Turtle(2015), the model is applied to verify the short-term and long-term effects of cumulative changes in investor sentiment.

$$R_t = a + bSum_{t-1,\Delta M+} + cSum^2{}_{t-1,\Delta M+} + \varepsilon_t \quad (7)$$

Following our above conclusions about the relationship between investor sentiment and stock returns, a positive estimate of parameter *b* is expected, because in short term, the increase of the sentiment will raise the demand of the stocks, which will increase the prices accordingly. Besides, a negative estimate of parameter *c* is expected, because for the long term, the value-return of the stock price will decrease the returns.

To verify the asymmetric effects of investor sentiment, we also consider Sum_{t,M_t-} as the sum of successive decreased sentiment during time period *t*.

$$Sum_{t,\Delta M_t-} = \begin{cases} Sum_{t-1,\Delta M_t-} + \Delta M_t, & \Delta M_t < 0 \\ 0, & \Delta M_t \ge 0 \end{cases} \quad (8)$$

Where $\Delta M_t = M_t - M_{t-1}$. For the initial time *t=0* in the sample, we set $Sum_{0,\Delta M_t-} = 0$. The model we use to test the effects of decreasing sentiment is similar with model (7).

DOI: 10.24818/18423264/52.1.18.11

$$R_t = a + dSum_{t-1,\Delta M-} + eSum^2_{t-1,\Delta M-} + \varepsilon_t \quad (9)$$

This time, the parameters *d* and *e* are expected to be insignificant. As we analyzed before, compared with short-selling, there is no such restriction for long side to hinder the value-return of the stock prices, and sentiment may have few effect on stock returns.

We use daily data of CSI 300 index and our investor sentiment index from November 1st, 2011 to September 30th, 2015 for analysis. The results are shown in Table 5.

Table 5. The effects of cumulative changes in investor sentiment

	b	*c*	*d*	*e*	*a*
formula(7)	0.014**	-0.013*	-	-	-0.001
	(2.43)	(-1.75)	-	-	(-1.38)
formula(9)	-	-	0.007	0.003	0.001
	-	-	(1.34)	(0.34)	(1.58)

The table reports t-value in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.

The results in Table 5 meet our expectations. The coefficient for $Sum_{t-1,\Delta M+}$ is positive and is significant at 5% level. The coefficient for $Sum^2_{t-1,\Delta M+}$ is negative and significant at 10% level. It shows that the short-term increased sentiment make the returns of the stock increased, and the medium-run increased sentiment decreased the returns. The coefficient for both $Sum_{t-1,\Delta M-}$ and $Sum^2_{t-1,\Delta M-}$ are insignificant. The results for robustness analyses not only verify our above conclusions about the relationship between investor sentiment and stock returns, but also prove the validity of our investor sentiment index from online stock forum.

5. Conclusions

This paper aims to study investor sentiment and its effects on stock returns. With the large-scale online stock forum data, we use a two-step text mining methods including data cleaning, text representation, feature extraction and two-step classification to identify individual investor sentiment, which is helpful for the accuracy of investor sentiment recognition, and for better test of relationship between investor sentiment and stock market returns. Besides, we propose a relatively comprehensive analysis to study the relationship between investor

DOI: 10.24818/18423264/52.1.18.11

sentiment and the stock market. First of all, analyze the prediction effect of individual investor sentiment on the stock market's returns. Second, study the different effects of investor sentiment on stock market in the short term and the medium run respectively. Finally, verify the mechanism through which investor sentiment affect stock returns.

We apply our sentiment recognition method to the most popular stock forum in China, and use 5,163,210 online posts on stock forum to compose the investor sentiment index. We test the effect of investor sentiment index on CSI 300 index from November 1st, 2011 to September 30th, 2015, and find individual investor sentiment from online stock forum has a strong effect on stock returns compared with institutional view. By establishing a 3 order VAR model, we verify a short-term positive effect and medium-run reverse effect of investor sentiment on stock market. By dividing sentiment into bullish stage and bearish stage, we verify the asymmetric effects of investor sentiment on stock market, and the bullish sentiment will get more obvious impact, which can also verify that investor sentiment influences stock returns by undervaluing or overvaluing stock prices. The robustness analysis with different time period and cumulative changes of investor sentiment also support our conclusions.

Our study not only provides the empirical case for behavioral finance, but also expands the application field of text mining technology. Besides, understanding the relationship between investor sentiment and stock returns is significant for both investors and supervisors.

Acknowledgement

This research was supported by National Natural Science Foundation of China (No.71771204, 71331005, 91546201), and the University of the Chinese Academy of Sciences(No.Y55202KY00).

REFERENCES

[1] **Antweiler, W., Frank, M. Z. (2004),*Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards;*** *Journal of Finance*, 59(3), 1259-1294;

[2] **Arvanitis, K., Bassiliades, N. (2017),*Real-time Investors' Sentiment Analysis from Newspaper Articles;*** *Advances in Combining Intelligent Methods*, 116, 1-23;

DOI: 10.24818/18423264/52.1.18.11

[3] **Baker, M., Wurgler, J., Yuan, Y. (2012),*Global, Local, and Contagious Investor Sentiment;*** *Journal of Financial Economics*, 104(2), 272-287;

[4] **Barberis, N., Xiong, W. (2010),*Realization Utility;*** *Journal of Financial Economics*, 104(2), 251-271;

[5] **Berger, D., Turtle, H. J. (2015),*Sentiment Bubbles;*** *Journal of Financial Markets*, 23, 59-74;

[6] **Black, F. (1986),** ***Noise;*** *Journal of Finance*, 41(3), 529-543;

[7] **Blum, A. L., Langley, P. (1997),*Selection of Relevant Features and Examples in Machine Learning;*** *Artificial Intelligence*, 97(1), 245-271;

[8] **Chau, F., Deesomsak, R., Koutmos, D. (2016),*Does Investor Sentiment Really Matter?*** *International Review of Financial Analysis*, 48, 221-232;

[9] **Cortes, C., Vapnik, V. (1995),** ***Support-vector Networks;*** *Machine Learning*,*20*(3), 273-297;

[10] **Das, S. R., Chen, M. Y. (2007),*Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web;*** *Management Science*, 53, 1375-1388;

[11] **Deng, N., Tian, Y., Zhang, C. (2012),*Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*;** Crc Press;

[12] **Fama, E., French, K., (1993),*Common Risk Factors in the Returns on Stocks and Bonds;*** *Journal of Financial Economics*, 33(93), 3-56;

[13] **Lee, C. M. C., Shleifer, A., Thaler, R. H. (1991),** ***Investor Sentiment and the Closed-End Fund Puzzle;*** *Journal of Finance*, 46(1), 75-109;

[14] **Lemmon, M., Portniaguina, E. (2006),** ***Consumer Confidence and Asset Prices: Some Empirical Evidence;*** *Review of Financial Studies*, 19(4), 1499-1529;

[15] **Long, J. B. D., Shleifer, A., Summers, L. H., Waldmann, R. J. (1991),*The Survival of Noise Traders in Financial Markets;*** *Journal of Business*, 64(1), 1-19;

[16] **Miller. E. M. (1977),*Risk, Uncertainty, and Divergence of Opinion;*** *Journal of Finance*, 32(4), 1151–1168;

[17] **Odean, T. (1998),*Are Investors Reluctant to Realize Their Losses;*** *Journal of Finance*, 53(5), 1775-1798;

[18] **Otoo, M. W. (1999),*Consumer Sentiment and the Stock Market;*** *Finance and Economics Discussion*;

[19] **Salton, G., Wong, A., Yang, C. S. (1975),*A Vector Space Model for Automatic Indexing;*** *Communications of the Acm*, 18(10), 613-620;

DOI: 10.24818/18423264/52.1.18.11

[20] **Scheinkman, J. A., Xiong, W. (2003),*Overconfidence and Speculative Bubbles;*** *Journal of Political Economy*, 111(6), 1183-1219;

[21] **Stambaugh, R. F., Yu, J., Yuan, Y. (2012),*The Short of It: Investor Sentiment and Anomalies;*** *Journal of Financial Economics*, 104(2), 288-302;

[22] **Stambaugh, R. F., Yu, J., Yuan, Y. (2015),*Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle;*** *Journal of Finance*, 70(5), 1903-1948;

[23] **Sun, L., Najand, M., Shen, J. (2016),*Stock Return Predictability and Investor Sentiment: A High-Frequency Perspective;*** *Journal of Banking & Finance*, 73, 147-164;

[24] **Tumarkin, R., Whitelaw, R. F. (2001),*News or Noise? Internet Postings and Stock Prices;*** *Financial Analysts Journal*, 57(3), 41-51;

[25] **Wheatley, S. M., Neal, R. (1998),*Do Measures of Investor Sentiment Predict Returns?*** *Journal of Financial and Quantitative Analysis*, 33, 523-547.

DOI: 10.24818/18423264/52.1.18.11