Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2010 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-1-2010

Can Web Forum Provide Valuable Information a Case Study on Chinese Stock Web Forum

Yang Hu

Xiangbin Yan

CAN WEB FORUM PROVIDE VALUABLE INFORMATION: A CASE STUDY ON CHINESE STOCK WEB FORUM

Yang Hu, Xiangbin Yan, Department of Management Science and Engineering, School of Management, Harbin Institute of Technology, Harbin, China.
E-mail: 313739803@qq.com xbyan@hit.edu.cn

Abstract

Web forum has become a major information sharing platform for the individual investor where may include much valuable information. In this paper, we apply automatic sentiment extraction methods to assess the public opinions on Chinese stock forms. A case study is conducted with Chinese stock Web forum. The case study shows that web forum can provide valuable information about the stock market and the discussion sentiment characteristic is significant correlated with stock market activities. Result of regression analysis shows the differences in investor sentiment has more power than sentiment index and message volume in prediction to stock index in China.

Introduction

With the widespread of the internet, many new social media appeared on the web, such as Web forums, Web blogs, etc. Web forum is one of the media widely used on the internet that people can publish opinions on various topics. In the financial areas, Web forum has become a major information sharing platform for individuals who have invested in stock market. Investors can share their views about stock market, investment experience, message about government policy, etc. People’s publishing behaviors on stock forums may reflect their invest opinions, prediction about stock market. Therefore, explore the activities of user in the stock forums can know their trading intentions and behaviors, which can provide a new approach to understand the stock market.

Though stock forum may include much valuable information about investment, however, assessing opinions on stock Web forum is a nontrivial task. The high volume of messages, the casual writing style, and the significant amount of noise require sophisticated text mining techniques to handle the data. In this paper, we apply text classification technology to analyze the publishing behaviors and sentiment trends of message board on a stock forum in China, and study the relationship among publishing behaviors, message sentiment and the stock market activity by correlation analysis and regression analysis. We find that there is correlation between stock forum and stock market.

The remainder of the paper is organized as follows. Section 2 reviews the background research on assessing and analyzing stock forum activity. Section 3 summarizes our research framework and methodology. In Section 4 we present the preliminary analysis of a case study on a Chinese stock forum, and discuss our findings comparing with related work. We conclude the paper with the extensions of this research in Section 5.

Literature Review

Investors use web forum as an information exchange platform can be traced back to the end of the last century. Many researchers and investors found some internet-related stocks’ performance cannot be explained by classical price theory, they began to seek other factors to explain this phenomenon [1]. Related literature are shown in Table 1, which illustrates authors, year, data employed, research methods, and conclusions.

Table 1 Literature Review

Author	Year	Data	Methodology	Conclusion
Wysocki[2]	1998	Over 3,000 stocks, Over 946,000 messages	Cross-sectional and Time-Series Analysis	Message volume can predict the next-day trading volume and return
Wysocki[3]	2000	55 firms with earnings-announcements	Regression	Announcement-period trading volume is positively associated with the level of announcement and pre-announcement message-posting activity
Tumarkin[4]	2001	73 Internet service stocks, 181,633 messages	Event study, VAR	No causal link exist between message posting and stock return/volume
Antweiler[5]	2002	313 stocks, 35 million messages	Regression	There are correlation between posting levels and stock volatility/return

Antweiler[6]	2004	45 stocks 1,559,621 messages	Contemporaneous Regression, NB,SVM	Message volume predicts trading volume and volatility
Das[7]	2005	4 stocks, over 170,000 messages	Text classification Correlation/Regression Analysis,	There is a close relation between sentiment levels and stock returns. Sentiment does not apparently predict returns
Das[8]	2007	24 tech-sector stocks, 145,110 messages	Regression, Text Classifier Algorithms,	Sentiment tracks index returns
Sabherwal[9]	2008	135stocks, over 160,000 Messages	Event Study, Lead-Lag Regressions	Message volume relate to abnormal return
Zimbra[10]	2009	Wal-Mart-related Web forum	Topic and Sentiment Analysis, Regressions	Topic and sentiment can predict stock price

The early research about stock forum is to investigate whether the message board contains valuable information, or it is just "noise". Recent studies have focused on correlation between the stock market and stock forum activities. With the development of complex text mining techniques, researchers are not limited to analyze message volume; they begin to study the content of messages, and try to identify the sentiment of the message, to investigate whether there is relation between message sentiments and stock trading activities.

The relationship between stock market and stock forum is a controversial issue, but recently more and more research evidences show that there are correlations. In China, scholars also pay attention to the stock market and sentiment, Wang et al. (2004) build a theoretical model and empirically prove that sentiment has a significant impact on the return and volatility in Shanghai and Shenzhen stock market[11]. Zhang et al. (2007) based on data from Shanghai and Shenzhen stock market, apply GARCH-M (1, 1) model to analyze the relation between Chinese stock market and investor sentiment. He found that the institutional investor’s sentiment has a significant influence on stock price, but the individual investor’s sentiment has no influence[12]. Liu et al. (2010) construct a new indictor of investment sentiment and study its impact on the return rates of portfolios using non-parameter statistics and regression models. Overall, investor sentiment will affect the stock that is vulnerable to effect by subjective factors in the valuation and volatility of sentiment has a cross-section effect[13]. Yang et al. (2010) investigate the different relations among investor sentiment, stock market returns and volatility in both bullish and bearish markets. Results show that investor sentiment does not influence returns in both states, but returns have significant impact on sentiment[14]. However, the measure of investor sentiment used in these researches is represented by proxy variables such as turnover, closed-end fund discount, RIPO, etc., and these proxy variables cannot reflect the emotional tendency of investors. To our best of knowledge, we have not found people conduct research on Chinese stock Web forum.

Methodology

In this study, we analyze the relationship between stock forum and stock market from two aspects of message volume and sentiment, the research framework is shown in Figure 1.

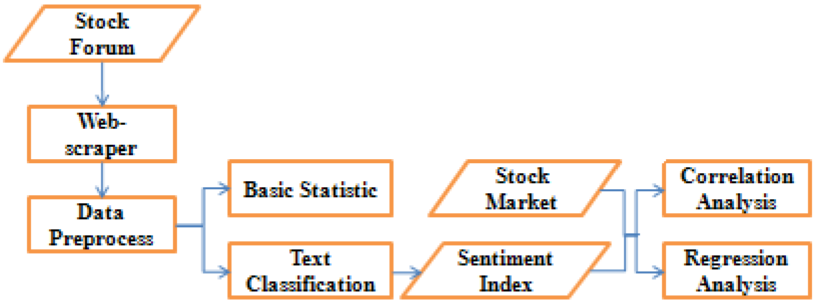

Figure 1 Research Framework

Data Collection

In the present work, we compare the Chinese major financial stock forums based on website traffic ranking and select number one as the data source. Then, we apply two components of MetaSeeker: MetaStudio and DataScraper to capture web page from Chinese Web forum. The main information we kept for each message includes dates, authors, and message content for further analysis.

Measure Employed

This paper aims to study the interaction between message in stock forum and the index changing in stock market, so we adopted similar measures which have been used in Das’s research for our analysis[8].

Message Volume (M), number of messages posted by users in the stock forum. We also calculate the message volume before and after the stock market opening. The quantity of messages before the opening is recorded as M1 and the quantity of messages during the opening is recorded as M2.

Sentiment Index (SI), aggregating the sentimental trend identified by text classifier. For messages on the daily message board, we use 1 represent bullish messages, -1 represent bearish messages, and 0 represent irrelevant messages. Similar to *Message Volume*, The index outside the stock market opening period is recorded as SI1, and the index during the opening period is recorded as SI2.

Disagreement (DIS), describe the differences in investor sentiment, calculated as follow, where R is the number of bullish messages and D is the number of bearish messages. This measure lies between zero (no disagreement) and one (high disagreement).

$$DIS = \left|1 - \left|\frac{R-D}{R+D}\right|\right| \qquad (1)$$

Stock Index (STI), we use Shanghai stock market closing index daily.

Volatility (VOL), because we want to examine the whole market, we select stock index instead of stock price used by Das (2007) to measure intra-day volatility as the difference between the high and low stock index for the day divided by the average of the open and closing index.

Trade Volume (TVOL), this is trading volume in number of shares traded in the day.

The former three are representatives of the stock forum, *and* the others are indicators of the stock market. All measures are aggregated daily and normalized by subtracting the mean value and dividing by the standard deviation of the data series.

Sentimental Classification

In this paper, messages are classified by text classifiers into one of three sentiment types: bullish, bearish and irrelevant. We use four text classifiers to process the message information including the Naive Bayesian, decision tree classifier, support vector machines, N-Gram, and finally the optimal classifier was selected as the classifier for sentimental identification.

Precision and recall are used to assess the four text classifiers [15]. Suppose that TP_i is the number of documents that belong to type C_i and are classified into C_i. FP_i is the number of documents that not belong to type C_i but is classified into C_i incorrectly. FN_i is the number of documents that ought to belong to type C_i but are assigned into other types. TN_i is the number of documents that not belong to type C_i and are not classified into C_i. Then the Recall and Precision of one classifier on the type C_i is as follow:

$$\text{Recall}_i = TP_i / (TP_i + FN_i) \qquad (2)$$

$$\text{Precision}_i = TP_i / (TP_i + FP_i) \qquad (3)$$

The above definitions are for a certain type C_i. To fully evaluate the classifier, we must consider all types. There are two ways to incorporate all types: Macro-Averaging and Micro-Averaging [16]. If C is the number of all types, the two definitions are:

Macro-Averaging:

$$MacAvg\_Recall = \frac{\sum_{i=1}^{|C|} \mathrm{Re}call_i}{|C|} \qquad (4)$$

$$MacAvg\_Precison = \frac{\sum_{i=1}^{|C|} \mathrm{Pr}ecision_i}{|C|}$$

(5)

Micro-Averaging:

$$MicAvg\_Recall = \frac{TP}{TP+FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i+FN_i)} \qquad (6)$$

$$MicAvg\_Precison = \frac{TP}{TP+FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i+FP_i)} \qquad (7)$$

Correlation Analysis and Regression Analysis

Based on the optimal classification results, we do correlation analysis on the measures mentioned above in order to study the direction and strength of correlations between stock forum and stock market. Because the measures we used are continuous variables, we choose Pearson correlation coefficient.

We run simple regression analysis to examine the prediction power of SI, M, and DIS to stock index. We focus on the effect of three measures on day t to t-5. We also construct an integrated model in which SI, M, and DIS are the independent variables. Therefore, our models at daily level are shown as following:

$$STI_t = b + \sum_{i=0}^{5} a_i SI_{t-i} \qquad (8)$$

$$STI_t = b + \sum_{i=0}^{5} a_i M_{t-i} \qquad (9)$$

$$STI_t = b + \sum_{i=0}^{5} a_i DIS_{t-i} \qquad (10)$$

$$STI_t = d + \sum_{i=0}^{5} a_i SI_{t-i} + \sum_{i=0}^{5} b_i M_{t-i} + \sum_{i=0}^{5} c_i DIS_{t-i} \qquad (11)$$

Experiment

Data

We conduct our experiment on a comprehensive Chinese stock web forum to evaluate our proposed framework. We chose Eastmoney (http://www.eastmoney.com/) as data source for our analysis. We choose it because of three reasons: First, it is one of the most well know Chinese stock web forums that have great influence on investors; Second, indicators such as web traffic and number of network users provided by Alexa official website (http://cn.alexa.com/) reveal that it ranks number one in all financial websites and ranks 61th of all the websites in China; Third, it is considered as an important information source for stock market, and has great influence on investor by experts in stock market. Because we aim to analyze the relationship between the whole stock market and stock forum, we choose the TanGuLunJin message board whose discussion focus on the whole stock market for our analysis, rather than single stock message board information.

For the stock board, the time span of the entire collected message is from Nov 1 2009 to Apr 30 2010, a total of 6 months, in which there are 123 trading days. We treat time 15:00(the stock market closing time) to 15:00(next day) as a day to aggregate the message. At the same time, we gather data separately before and after 9:30(the stock market opening time) in order to analyze the effect of message board in different periods. Totally, we get 17062 messages, 14529 messages in trading days. The maximum message volume is 450 per day, and the minimum is 20 per day. There are more messages in the trading days, specifically, 95 message per day on average, and 118 per day on average for the trading days. Figure 2 shows the distribution of messages per trading days. The message volume is relatively stable except on 30, April the day may be effected by the following International Labor Day.

We aggregate messages by hour during the trading days to examine publishing patterns of participants, which is shown in Figure 3. We can see the messages distribute from 7 o’clock to 24 o’clock and at 15 o’clock reach the highest point.

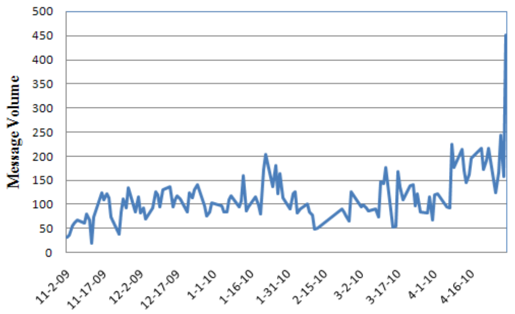

Figure 2 Distribution of Message Volume Daily

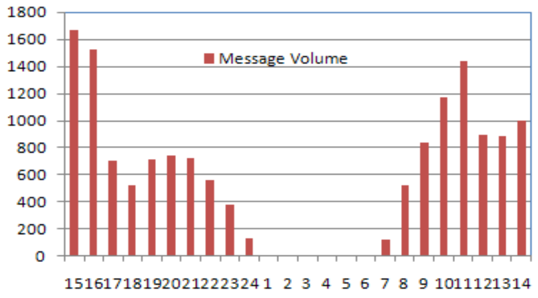

Figure 3 Distribution of Message Volume by Hour

Sentimental Identification

We use text classification methods to classify messages to identify their sentimental trend. Four classifiers are used, which namely Support Vector Machine (SVM), Decision Tree, NaiveBayes and N-Gram. We use recall and precision mentioned above to choose the best classifier. Basic text classification process is shown in Figure 4.

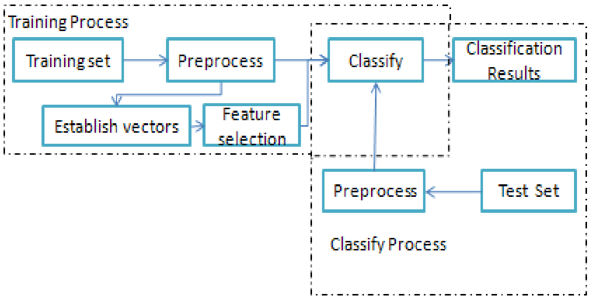

Figure 4 Basic Text Classification Process

Training set: we extraction 500 messages randomly from the whole data set during trading days and classify them manually to identify their sentimental trends. Each message is classified into one of the three types of bullish, bearish, and irrelevant. In the training set the number of messages belongs to the three types are 128, 104, and 268 respectively.

Preprocess: we use ICTCLAS (Chinese Lexical Analysis System, http://ictclas.org/) as a tool for Chinese word tagging. All messages are segmented to words with tag for feature selection.

Establish vectors: we need to transfer text data into a vector form by VSM (vector space model) when using SVM, NB and Decision Tree. We first construct a lexicon including all terms in the whole data sample. Each term is considered as an attribute of the document. In this paper, the document is referred to messages, and we get a total of 5683 attributes. Then we transfer the documents into a collection of attributes value. The number of times the term appears in one document is the corresponding value of the attribute and a series of attribute value constitutes a vector form of the document.

Feature selection: When the number of features is large, feature selection may improve the classification performance by selecting an optimal subset of features. Previous classification studies usually include some form of feature selection in order to extract the most important words or phrases. In this study, we use Information Gain (IG) heuristic to conduct feature selection, thus building selected feature sets. All features with an information gain greater than 0.0025 are selected. In the end, we get 105 attributes.

Classification: we use two kinds of software (WEKA and Lingpipe) to conduct sentimental classification. WEKA (Waikato Environment for Knowledge Analysis) is data mining software, which provides methods including SVM, Decision Tree (J48), and NaiveBayes. Lingpipe is a natural language processing package which provides N-Gram classification method.

A preliminary classification result is shown in Table 2. For the four classifiers, NaiveBayes and SVM are better. But the result was not good enough for use. After weed out irrelevant data, we classify the data into two types (bullish and bearish) and run the classifiers again. The result is shown in Table 3. The recall and precision of all four methods has increased, which indicates the existence of irrelevant data has side effect on the classification results.

Table 2 Classification Result (Three Types for All Data)

		N-Gram	NaiveBayes	J48	SVM
Recall	MacAvg	0.488	0.541	0.455	0.490
	MicAvg	0.552	0.634	0.554	0.630
Precision	MacAvg	0.504	0.587	0.472	0.653
	MicAvg	0.552	0.634	0.554	0.630

Table 3 Classification Result (Two Types for Bullish and Bearish Messages)

		N-Gram	NaiveBayes	J48	SVM
Recall	MacAvg	0.600	0.747	0.626	0.747
	MicAvg	0.589	0.634	0.554	0.630
Precision	MacAvg	0.613	0.758	0.625	0.758
	MicAvg	0.589	0.634	0.554	0.630

To remove the impact of the irrelevant data, we first divide training set into two categories, relevant (bullish and bearish) and irrelevant, 232 and 268 messages respectively. Then we classify the data with four classifiers. Table 4 shows the result. Compared to other classifiers, SVM classifier gets the best result. After removing irrelevant data, we re-use the above four classifier to process the related data. In the first process, some data belong to irrelevant is classified into related, so in this step, we still consider data as three types (bullish, bearish and irrelevant). Experimental result is presented in Table 5. Recall and precision of SVM classifier ratio were significantly higher than the others, while decision tree classifier is the worst.

Table 4 Classification Result (Two Types for All Data)

		N-Gram	NaiveBayes	J48	SVM
Recall	MacAvg	0.603	0.672	0.651	0.658
	MicAvg	0.626	0.656	0.656	0.678
Precision	MacAvg	0.632	0.705	0.653	0.743
	MicAvg	0.626	0.656	0.656	0.678

Table 5 Classification Result (Three Types for Relevant Data Get from Table 4)

		N-Gram	NaiveBayes	J48	SVM
Recall	MacAvg	0.513	0.531	0.342	0.648
	MicAvg	0.518	0.571	0.351	0.762
Precision	MacAvg	0.504	0.578	0.338	0.896
	MicAvg	0.518	0.571	0.351	0.762

In summary, SVM classifier is the best of the four and we choose it as final method to undertake sentimental classification. In the classification process, we first weed out the irrelevant by classifier like Table 4 to improve the accuracy for future classification. Then we classify the reminder data using SVM.

Correlation Analysis between Stock Forum and Stock Market

Message Volume and Stock Market

Correlation analysis is conducted using statistical software SPSS, the result is shown in Table 6. At the 0.01 significance level, message volume has negative correlation with daily stock index. We further study correlation during sub-periods, and found that message volume has negative correlation with daily stock index no matter in which sub-periods. At the 0.05 significance level, message volume during the opening period has negative correlation with trading volume. However, there is no significant correlation between message volume and stock market volatility at both periods.

Table 6 Correlation between Message Volume and Stock Market

		STI	TVOL	VOL
M	Pearson Correlation	-.307**	-.161	.090
	Sig. (2-tailed)	.001	.075	.320
	N	123	123	123
M1	Pearson Correlation	-.259**	-.108	.019
	Sig. (2-tailed)	.004	.234	.832
	N	123	123	123
M2	Pearson Correlation	-.306**	-.192*	.155
	Sig. (2-tailed)	.001	.033	.088
	N	123	123	123
** Correlation is significant at the 0.01 level * Correlation is significant at the 0.05 level				

Sentiment Index and Stock Market

There are 14,529 messages altogether in 123 trading days. We got 3201 valid messages after the classification preprocesses, in which 752 messages are with bullish sentiment, 1223 messages are with bearish sentiment, and 1226 are irrelevant. Figure 5 shows the trend of standardization sentiment index and stock index during trading days. It can be seen that the trend of SI and STI are similar to each other and there may be a correlation. Figure 6 shows frequency distribution of the sentiment index during the trading days. Y-axis refers to number of the days which get the sentiment index. We can see that sentiment index for the majority of trading days is positive and distributes from 0 to 9 points with average sentiment index 3.97 and the median number is 5.0. Frequency distribution of the sentiment index shows that attitude of investors to the stock is positive during majority trading days, but it’s not in contradiction with conclusions mentioned above that most of the valid messages are with negative sentiment. We attribute this phenomenon to that most negative information is concentrated on certain days. For example, on the day April 30, 2010, there are 51 negative messages and sentiment index fell to -32.

Correlation analysis results between stock market activities and sentiment index of message boards in different period are shown in Table 7. Sentiment index of each period is positively correlated with stock index and trading volume. DIS is negatively correlated with stock index and trading volume. It suggests that the greater the difference in investors’ sentiment, the worse performance of the stock market. At the same time, all measures are not correlated with the volatility.

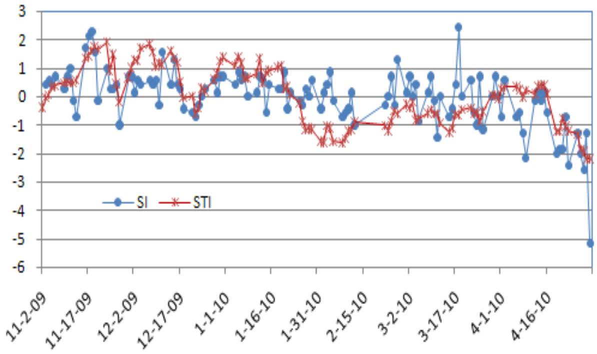

Figure 5 Trends of Sentiment Index and Stock Index

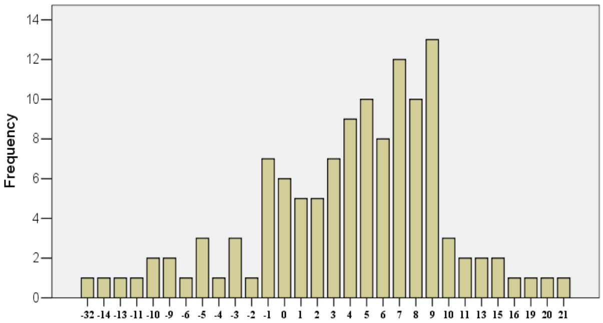

Figure 6 Frequency Distribution of the Sentiment Index

Table 7 Correlations between Sentiment Index and Stock Market

		STI	TVOL	VOL
SI	Pearson Correlation	.555**	.330**	-.068
	Sig. (2-tailed)	.000	.000	.456
	N	122	122	122
SI1	Pearson Correlation	.545**	.351**	-.072
	Sig. (2-tailed)	.000	.000	.430
	N	121	121	121
SI2	Pearson Correlation	.411**	.208*	-.028
	Sig. (2-tailed)	.000	.028	.770
	N	112	112	112
Dis	Pearson Correlation	-.588**	-.387**	.041
	Sig. (2-tailed)	.000	.000	.651
	N	122	122	122
** Correlation is significant at the 0.01 level * Correlation is significant at the 0.05 level				

Regression Analysis

Performance of regression models is shown in Table 8. Obviously, all four models are significant and DIS has more prediction power than the other two. That is consistent with the result of correlation analysis that DIS is the most related to stock market. In the integrated model, the overall fit of the model is increased to an R^2 value of 0.600. The increase suggests that the combination of the three measures can enrich the predictive power and provide information more closely with stock market.

Table 8 Performance of Regression Models

	Model summary	ANOVA	
Model	R^2	F	Sig
8(SI)	0.452	15.146	.000
9(M)	0.214	5.033	.000
10(DIS)	0.548	22.252	.000
11(All)	0.600	8.157	.000

Comparison with related research

Based on our data sample, the present work has proven that there is much valuable information in Chinese stock forum and significant correlation between stock forum and stock market, which is consistent with the conclusions of other researches [6] [8]. However, our results also show some difference. First, there is no correlation between volatility and measures of the stock forum in China, which was found significant in other research [5]. Second, in China the correlation among Disagreement, stock index and trading volume is stronger which is not consistent with the findings of Das[8], who find that the disagreement measure evidences very little correlation to other variables. Third, the result of regression analysis shows that Disagreement has the best predictive power to stock index which is not found in other researches. Furthermore, for our data sample, the integrated model combining SI, M and DIS has more power in prediction (greater value of R^2) than models constructed by Zimbra[10] and Sabherwal[9].

Conclusion

In our research, a variety of text mining and classification methods are used to analyze the sentiment of message board information in Chinese stock forum message board. Each message is classified into bullish, bearish or irrelevant. We examine and analyze the correlations between stock forum activities and stock market activities from two aspects of message volume and sentiment. We get the following conclusions: First, message volume and the stock index have a negative correlation and is negative correlated with trading volume during the stock market opening. Second, sentiment index is positive correlated with stock index and trading volume. Third, Disagreement is negatively correlated with stock index and trading volume. Fourth, all measures have no correlations with volatility of stock market. Finally, in the regression analysis, we find that the disagreement (DIS) has more power to predict stock index and the combination of sentiment index, message volume and disagreement can improve the prediction power. The overall evidence suggests that, in the Chinese stock market, the information in the stock forum has some correlations with stock market activities but the correlation pattern is different from that abroad.

The present work is a preliminary study in Chinese stock Web forum. Future research can refer to the following aspects: developing more efficient text classification algorithm for content on stock forum; based on the significant correlations, study can focus on analyzing the causality between stock forum and stock market and setting up the appropriate forecasting model; starting from the level of individual stocks, study the effect of stock forum on stock market.

References

[1] Zhang, Y. and Swanson, P. Are day traders bias free?evidence from internet stock message boards. *Journal of Economics and Finance*, 2010,34(1):96-112.

[2] Wysocki, P. Cheap talk on the web: The determinants of postings on stock message boards.Working paper, Univercity of Michigan,1998.

[3] Wysocki, P. Private Information, Earnings Announcements and Trading Volume, or Stock Chat on the Internet. A public Debate about Private Information, Working Paper, University of Michigan Business School. *Ann Arbor*, 2000, 48109-41234.

[4] Tumarkin, R. and Whitelaw, R. News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 2001,57(3):41-51.

[5] Antweiler, W. and Frank, M. Internet stock message boards and stock returns. *University of British Columbia, Working Paper*,2002.

[6] Antweiler, W. and Frank, M. Z. Is all that talk just noise? The information content of Internet stock message boards. The Journal of Finance,2004,59(3):1259-1294.

[7] Das, S., Martinez-Jerez, A. and Tufano, P. eInformation: A clinical study of investor discussion and sentiment. *Financial Management*, 2005,34(3): 103-137.

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

[9] Sabherwal, S., Sarkar, S. and Zhang, Y. Online talk: does it matter? *Managerial Finance*, 2008,34(6): 423-436.

[10] Zimbra, D, Fu, T and Li, X. Assessing public opinions through Web 2.0: A case study on Wal-Mart. *The 13th International Conference on Information Systems*,2009.

[11] Meijin, W. and Jianjun, S. Stock Market Returns, Volatility and the Role of Investor Sentiment in China. *Economic Research Journal*, 2004,10:75-83.

[12] ZhANG Qiang, YANG Shu-e, YANG Hong.An Empirical Study on Investors' Sentiment and Stock Returns in Chinese Stock Market.*Systems Engineering*,2007,07:13-17.

[13] LIU Li-ya, DING Jian-ping,CHEN Zhen-yu, XIANG Heng-ning.Empirical Study on Influence of Investment Sentiment on the Stability of Capital Market:Analysis of Cross-sectional Effect. *Journal of Finance and Economics*, 2010, 03:133-143.

[14] YANG Yang,WAN Di-fang. Relationship among Investor Sentiment,Stock Market Return and Volatility in Different Market States. *Systems Engineering*, 2010, 1(28):19-23.

[15] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 1998: 137-142.

[16] Mitchell, T.M. Machine learning. 1997, New York: McGraw-Hill Co., Inc.

Acknowledgements

This work was supported by National Natural Science Foundation of China under the number of 70601009 and 70890082.