Baidu Index sentiment and its impact on Chinese stock price volatility[1]

ABSTRACT

Nowadays, search engine use increasingly reflects investor sentiment, which affects the return on the stock market. In this article, we examine the relationship between Baidu Index sentiment and China's stock market returns. In two different GARCH models, the benchmark model and a Baidu Index extended model, the one-step forward method is used to predict the return of stock market. The study finds that Baidu Index − search volume is a valid indicator for forecasting volatility in China's stock market. The Baidu Index extended model performs better than the benchmark model, both in periods of high volatility and periods of low volatility. These results are quite robust in Shanghai Stock Exchange, Shenzhen Stock Exchange, CSI 300 Index, and CSI 100 Index. This study shows that the investor sentiment reflected in the Baidu Index can be used as a good early warning indicator of China's stock market.

1. Introduction

In the stock market, reliable information is an intangible asset. Especially in times of financial turmoil, investors want to get relevant information as soon as possible in order to make fast decisions. Search engines offer an easy and low-cost means for investor to access information. Investor sentiment, which is reflected in search terms and search volume, can be considered as an indicator for stock market forecasting. In China, Baidu as the leading search engine provides investors a convenient channel to access relevant information. Data about search terms and search volumes can be obtained through Baidu Index. These investor sentiment data may reveal the effects of market shocks on the stock market.

Preis 2010, Bordino 2012 & 2014; Vlastakis 2012, Dimpfl and Jank 2016, and Caporin, 2017 examined the relationship between the number of daily queries and trading volume and volatility for a given stock. The results suggest that there is a circular relationship between the search volume of Google Trends and the stock volume, returns, liquidity and volatility. Autoregressive models with search queries are shown to perform better than traditional time series model for both in-sample and out-of–sample testing. Particularly, in the period of high volatility their forecasting accuracy increases with the number of investors' search queries. Mao 2011 and Ackert 2016 compare online data such as Twitter feeds, news headlines, and search engine queries with traditional investor sentiment survey data in the predictive validity. They conclude that the traditional investor sentiment survey data are only a lagging indicator of financial markets and the Google Trends search volume online data etc. have better predictive capability. Da (2011) finds that based on the Russell 3000 stock data an increase in Google queries will result in stock price increase for the following two weeks. However, Preis

[1] **Acknowledgments:** the authors thank anonymous reviewers, editors for their helpful comments and advice. We also thank to the financial supports from The National Social Science Fund of China (16BJY052).

(2013) finds that an increase in certain search terms, such as debt, is often related to stock selling. Pyo (2017) finds that stock market return is negatively correlated with search frequency over the same period. Bank 2011, Zhang 2013 and Da 2015 find that high search volume is related to high returns in the short run but the opposite in the long run. Some studies examine the relationship between search volume and firm-level investment return. Kristoufek 2013 and Rochdi 2015 argue that the Google Trends search term helps to increase return on investment and diversify portfolio risk. Heiberger (2015) finds that at macro-level public concern about bad news regarding large corporations could serve as an early warning signal for systemic risk in the financial sector during financial crisis. In the case of developing countries such as Brazil and Indonesia, Ramos 2017 and Usman 2017 find that there might be an inverse relationship between financial markets and Google's search volume, that is financial market impacting Google search queries. By contrast, some studies do not find significant improvement in prediction by models that include search volume. Curme (2014) argues that the decline in the predictive power of search terms may reflect the increasing integration of Internet data into automated trading strategies.

Most of the existing studies use Google Trends to study the relationship between search volume and stock market volatility in developed country markets such as the United States. However, study on the impact of investor sentiment on stock markets in emerging economies using non-Google data is still relatively rare. The contribution of this article is using big data from Baidu, China's leading search engine, to examine whether a model with Baidu Index sentiments included can improve the prediction of Chinese stock market return. To test the robustness of the model, we will also examine the model in Shanghai Stock Exchange, Shenzhen Stock Exchange, CSI 300 Index and CSI 100 Index.

The reminder of the article is structured as follows. Section 2 describes data collection and search term selection. Section 3 discusses the methodology. Sections 4 conducts estimation and forecasting. Section 5 presents empirical results. Section 6 concludes.

2. Data collection and keywords selection

The closing prices of different Chinese stocks in this article are all from Wind Info. The formula for calculating the average daily return time series of stock market is as follows:

$$Return_{i,t} = \frac{(Closing\ \ price_{i,t} - Closing\ \ price_{i,t-1})}{Closing\ \ price_{i,t-1}} \tag{1}$$

In the above formula, $Closing\ \ price_{i,t}$ represents the closing price of stock market i on the t day.

Google was ranked first in the Chinese search engine market before its leaving China in 2010. Since then, Baidu’s market share rose from 60.2% in 2010 to 77.9% in 2017 (Statcounter, 2017). Given Baidu’s leading position in Chinese search engine market, we use Baidu Index instead of Google Trends for source of search data (http://zhishu.baidu.com/) for investor sentiment .

Baidu Index provides us with time-series data for each investor sentiment keyword. Although Baidu Index does not provide query result data, we use web crawler to obtain

numerical values. In the Chinese stock market, the most commonly used keywords by investors for market trend are "Niushi (Bull market)" and "Xiongshi (Bear market)". We also included other keywords that influence the price movement into the model, such as Fangjia (Housing prices), Tongzhanglv (Inflation rate), Shiyelv (Unemployment rate), etc., but none of them is significant. Hence, we excluded these keywords.

Before empirical analysis, we need to examine the descriptive characteristics of the data. J-B statistics examine skewness and kurtosis for time series data. The null hypothesis is normal distribution. Table 1 shows that Shanghai stock market returns and search data are not subject to the normal distribution. At the 1% significance level, the return data is negatively biased while the search index is positively biased. The kurtosis is convex for of all the data.

Table 1

The descriptive statistics of the Shanghai stock return and Baidu Index.

	Mean	Median	Max	Min	Std.Dev	Skew.	Kurt.	J-B
SHRET	0.020	0.062	5.759	-8.488	1.402	-0.884	9.273	2934.803***
LNNIUSHI	6.081	5.965	8.395	5.215	0.488	1.739	6.215	1549.200***
LNXIONGSHI	5.803	5.697	8.516	5.043	0.431	2.445	10.705	5753.413***

***indicates 1% significance level.

Before model prediction, we use the unit root test to examine stability for the data. The null hypothesis is that there is a unit root. The test results show there exists a unit root in the time series of the closing stock prices. However, the difference in stock market rate takes care of this issue. There does not exist unit root in the natural logarithm of search volume data. Table 2 shows the unit root test results on the natural logarithms of returns on the Shanghai stock market and Baidu search data. As can be seen from Table 2, at 1% significance level, all the time series in the sample period are stationary.

Table 2

Unit root tests on the Shanghai stock return and Baidu Index.

	ADF	PP
SHRET	-38.80500***	-38.78309***
LNNIUSHI	-3.932681**	-5.497832***
LNXIONGSHI	-6.379709***	-11.62937***

***indicates 1% significance level.

3. Methodology

In financial time series models, the stability of disturbance variance is usually worse than it is assumed. As a result, prediction errors take place in groups. That is, there is heteroscedasticity. The conditional variance of the error term usually varies with time and depends on the magnitude of the past error. Engle (1982) proposed the ARCH model accordingly. The idea is that the conditional variance (σ_t^2) of the perturbation term u_t depends on the residual difference squared u_{t-1}^2 of the previous moment. A p-order

autoregressive conditional heteroscedastic ARCH (p) model is:

$$y_t = x_t\gamma + u_t, \qquad t=1,2,...,T \tag{2}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i u_{t-i}^2 \tag{3}$$

Using Ljung-Box Q residual square correlation diagram and ARCH LM test respectively, we investigate whether the ARCH model contains residual effects for Shanghai Stock Exchange.

First, we use Akaike Information Criterion to determine that the mean equation follows the AR (2) MA (2) process. Then we use Ljung-Box Q to test residual square autocorrelation and partial correlation for the mean equation. Finally, the null hypothesis is that the data are independently distributed. That is, the overall correlation coefficient is zero, and any observed correlations are the result of error in random sampling only. We use 10 lags to calculate the results in Figure 1. The coefficients are not 0 significantly, and Q statistic is significant, indicating that residuals of formula (2) have significant ARCH effect.

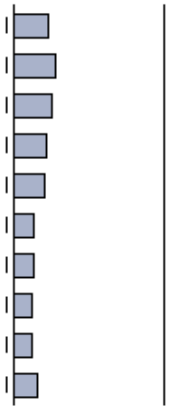
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.233	0.233	90.273	0.000
		2	0.289	0.248	228.72	0.000
		3	0.266	0.177	346.20	0.000
		4	0.223	0.098	428.97	0.000
		5	0.217	0.082	507.55	0.000
		6	0.132	-0.018	536.65	0.000
		7	0.144	0.014	571.29	0.000
		8	0.129	0.020	598.83	0.000
		9	0.124	0.030	624.33	0.000
		10	0.170	0.091	672.47	0.000

Fig.1. Autocorrelation and partial correlation diagrams of the residual square of the mean equation

The fluctuation of residual squared is shown in Figure 2. Residual squared is shown to have significant time-varying and clustering characteristics. Hence, GARCH type models are appropriate to use.

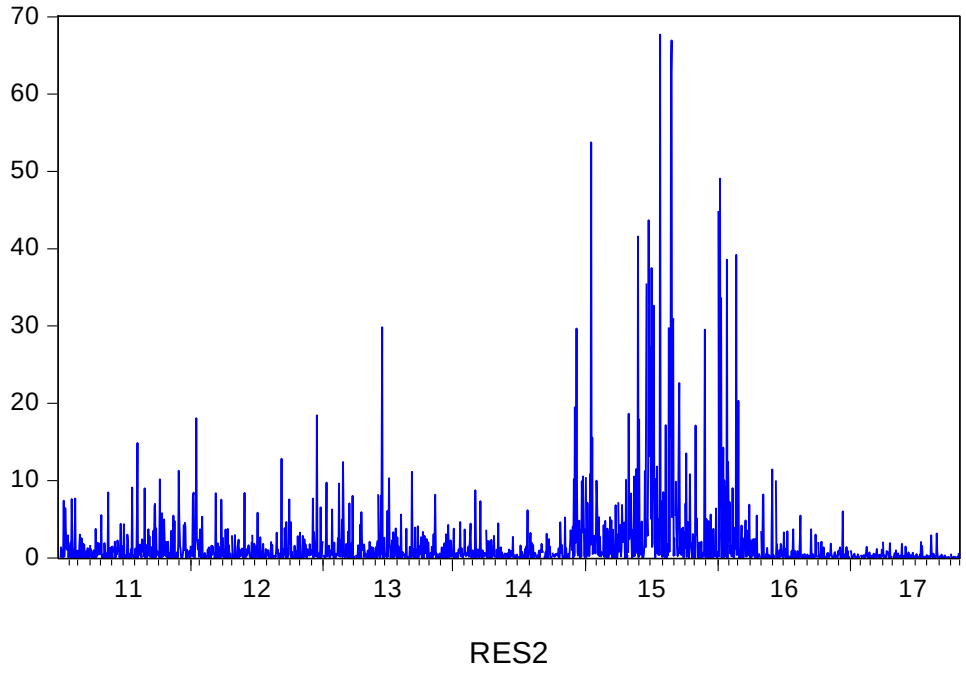


Fig. 2. Residual square line graph

The ARCH LM test is calculated by an auxiliary test regression. To test the null hypothesis that there is no ARCH effect for residual series up to the p-order, the following regression is needed:

$$\hat{u}_t^2 = \beta_0 + (\sum_{s=1}^{p} \beta_s u_{t-s}^2) + \varepsilon_t \tag{4}$$

$\hat{u}_t$ is the residual. Equation (4) represents the regression of residual squared $\hat{u}_t^2$ over a constant and the sum of residual squared up to pth order lag. The test regressions include F-test and $T \times R^2$, where T is the number of samples for assistance regression. LM test usually follows $\chi^2(p)$ distribution.

Table 3
LM test.

Heteroskedasticity Test: ARCH			
F-statistic	104.6688	Prob. F(2,1650)	0.0000
Obs*R-squared	186.1066	Prob. Chi-Square(2)	0.0000

We do conditional heteroscedasticity ARCH LM test for equation (2). The results of ARCH LM test for lag order p = 2 are shown in Table 3. It show that the Prob value is 0, rejecting the null hypothesis and indicating that there is an ARCH effect in the residual sequence of equation (2). Hence, we can use GARCH model for further study.

4. Estimation and Forecasting

The GARCH model was proposed by Bollerslev (1986). The general form of a GARCH (p, q) model is:

$$\sigma_t^2 = \omega + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + \sum_{i=1}^{p} \alpha_i u_{t-i}^2 \tag{5}$$

Where $\omega > 0$, $\beta_j \geq 0, j=1,...,q$, $\alpha_i \geq 0, i = 1,...,p.$

Model settings are more common in the financial sector. Traders tend to predict current variance by a constant variance, a priori variance forecast, and a weighted average of the previous innovation. If the rate of return on assets rises or falls a lot, traders will increase their estimates of the next-period variance. This form of variance can reveal a clear fluctuation cluster phenomenon in the return on financial assets. Current high returns are more likely to be accompanied by higher returns or losses, which means there is a greater likelihood of a large loss when the return on assets is high.

In order to estimate the exact values of p and q, we estimate the different GARCH (p, q) and judge it according to the AIC value of the variance equation. The results show that we select GARCH (1,1) is more appropriate. Therefore, the variance equation (5) can be simplified as a reference model:

$$\sigma^2 = \omega + \beta_1 \sigma_{t-1}^2 + \alpha_1 u_{t-1}^2 \tag{6}$$

This is the basic model we use in Baidu Index data extended model. The optimal lag term for the extended model with Baidu Index data remains to be determined. We estimate hysteresis from 0 to 4 . At a 5% level of significance, a lag of 0 has a significant effect on the variance. Thus, due to Baidu Index search with real-time, the current investor sentiment will have instant impact on the today's return on the stock market. There is no lag between the two variables. Therefore, we choose the following conditional variance model. The extended model that contains the Baidu Index is:

$$\sigma^2 = \omega + \beta_1 \sigma_{t-1}^2 + \alpha_1 u_{t-1}^2 + \lambda LnBaiduindex \tag{7}$$

Based on the above model, we take a step forward to predict the return of SSE A shares. Our sample period is from January 4, 2011 to October 31, 2017. First, we use the data estimation model from January 4, 2011 to December 30, 2011. Then use this information to predict the January 4, 2012 variance. And then repeat the process, through a one-step forward method to predict the variance of the entire sample period. Then assess the effect of out-of-sample estimates after the end of the estimation period. By the end of the sample period until October 31, 2017. This ensures that each observation period is examined, and then the prediction error can be used to determine the accuracy of the prediction.

5. Empirical Findings

The final prediction error value, whether it is a basic model or an extended model that contains the Baidu Index, is the difference between the model's actual and predicted conditional variance. The forecast results are shown in Figure 3. Overall, the Baidu Index extended model has a smaller prediction error than the basic model, indicating that

incorporating search data into the model can improve forecasting. In addition, an interesting observation is that in high volatility period from October 2012 to May 2013 and from mid-November 2014 to mid-May 2016, neither the basic model nor the extended model did not well explain its variance. The prediction errors for both models were significantly higher than normal one during this period. However, there are some differences between two models. During high volatility, the error between the predicted and actual values of the extended model containing Baidu Index is significantly smaller than that of the basic model. This shows that the extended model containing Baidu Index can predict the change of volatility more accurately than the usual GARCH model. From this we can conclude that in the market turbulence period, as investors seek more market information to avoid risks, resulting in increased search volume and sharp fluctuations in the return on the stock market.

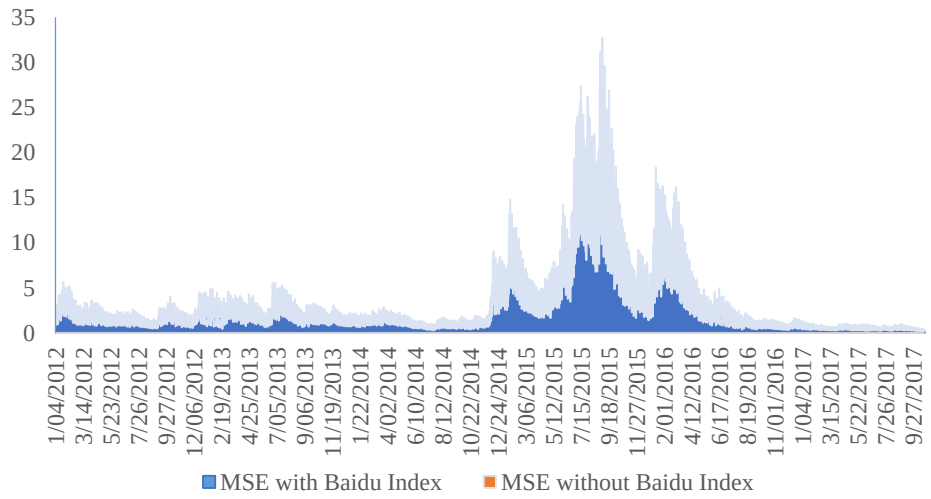


Fig. 3. MSE differences over the sample period between two models

In order to accurately judge the extent to which the extended model relative to the basic model improves the prediction, we calculated the root mean square error of two models. In addition, we also calculated the U1 parameter, which examines the accuracy of the prediction. The value is between 0 and 1. The closer to 0, the higher the prediction accuracy is. The root mean square error and U1 parameter improvement degree are shown in Table 4. The second and third rows of the table show two models, namely, with and without the Baidu Index, respective MSE and U1 parameters, and the degree to which the percentage is decreasing. The results show that the Baidu Index extended model can reduce the MSE by 3.07% and reduce the U1 parameter by 15.24 % compared with the basic model without Baidu Index.

Table 4
Predictive comparison.

Models	U1 Theil	U1 Theil reduction	MSE	MSE reduction
Basic Model	0.950591		0.922881	
Extended Model with Baidu Index	0.805726	15.24%	0.894504	3.07%

In order to examine the differences in predictive quality improvement over time, we classify the sample into different time periods. First, we divide the subsamples into high and low volatility based on the level of volatility. High volatility refers to the period when the conditional variance exceeds the mean. Low volatility means a period below the mean. Its division is shown in Figure 4.

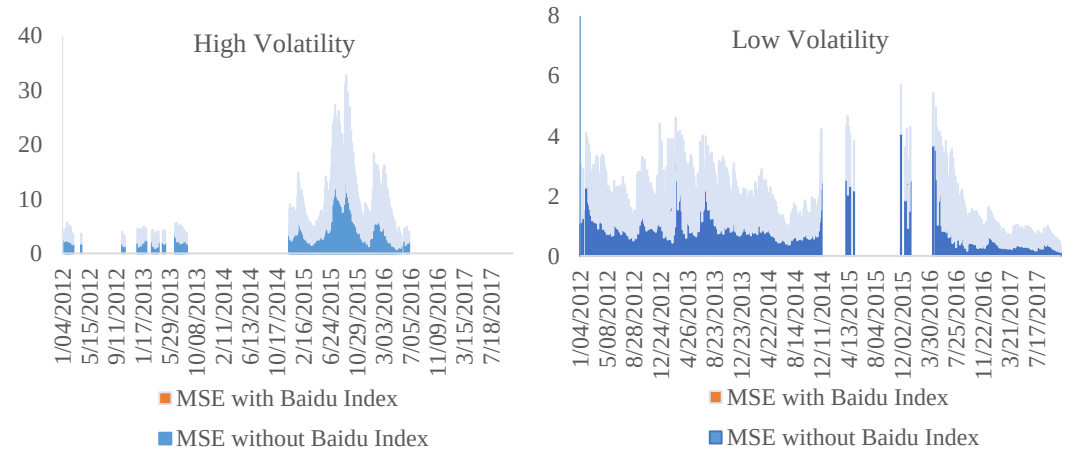
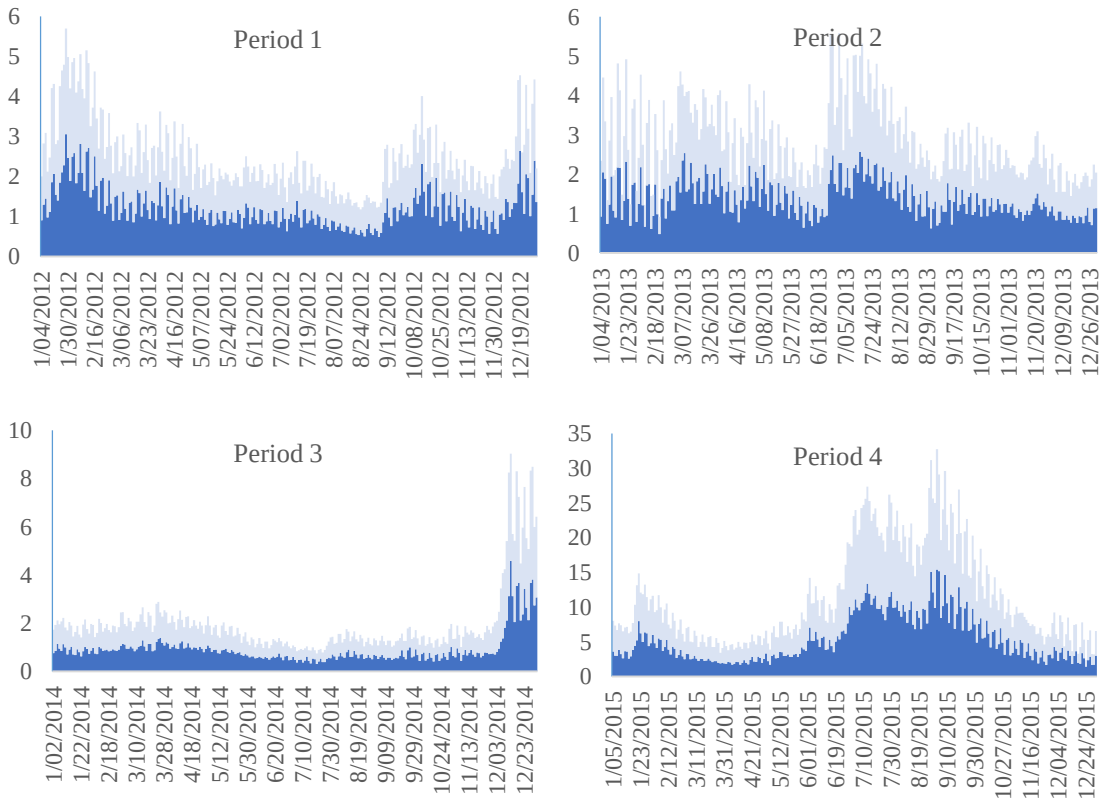


Fig. 4. MSE differences during periods of high and low volatility

In addition, the forecast results are also evenly divided into 6 different time periods. Each year is a period of time. For example, Period 1 is from January 4, 2012 and ends December 31, 2012. Period 2 is 2013, and so on. Figure 5 shows the MSE differences between the two prediction models in different periods.



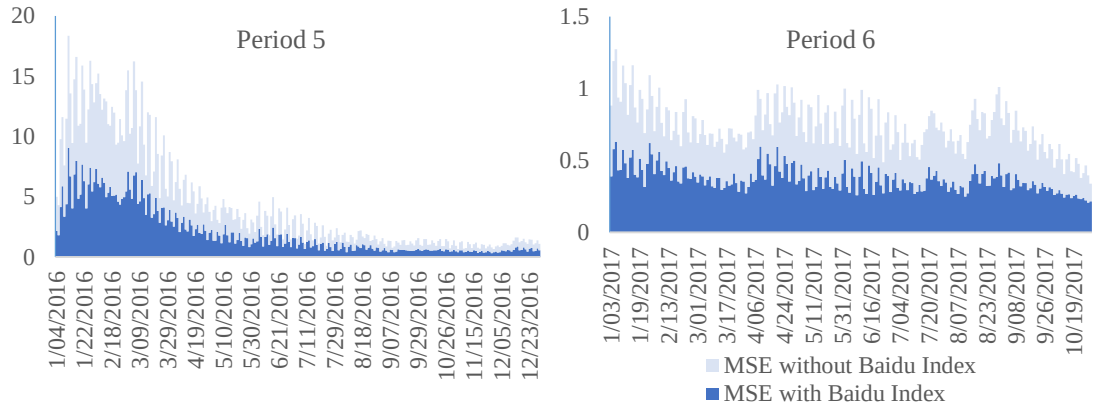


Fig. 5. Comparison of prediction accuracy of two models in different periods

Table 5 shows that the extended model containing the Baidu Index improves the accuracy of the prediction over time. Theil and MSE dropped 15.5% and 3.1%, respectively, even during periods of low volatility . Relatively high volatility, 2014 years (Period 3) and, 2015 years (Period 4) MSE degree of decrease was significantly better than 2012 year, etc. of low volatility. As a whole, the extended model containing the Baidu Index has significantly better predictive ability than the basic model at any time.

Table 5
Prediction differences in different periods.

Baidu Index vs baseline	U1 Theil reduction (%)	MSE reduction (%)
High volatility	22.8	8.1
Low volatility	15.5	3.1
Period 1	12.6	0.3
Period 2	16.6	1.9
Period 3	21.5	7.2
Period 4	16.4	5.2
Period 5	15.3	5.2
Period 6	18.6	0.3

In order to investigate the robustness of the benchmark and extension models in predicting the performance of different markets. We also examine Shenzhen A shares, CSI 300 Index and CSI 100 Index. The results of the robustness test are shown in Table 6.

Table 6
Robustness tests in different stock markets.

Baidu Index vs baseline	U1 Theil reduction (%)	MSE reduction (%)
Shenzhen	19.2	3.2
CSI 300	15.1	2.1
CSI 100	13.7	1.7

The results show that Baidu Index search data is suitable for stock market forecast. In addition, the extended model containing Baidu Index in the high volatility period performed

better than the low volatility period. As a result, investors can use the Baidu Index as an early warning indicator to predict future market trends.

This helps to deal with turmoil in the stock market in the early stages. So you can use the appropriate hedging tools to reduce market risk.

6. Conclusion

At present, the use of Baidu Index's search data to predict the volatility of returns in different stock markets in China is still relatively rare. This article focuses on investor sentiment and the volatility of the return on the stock market. Therefore, the use of search engines to retrieve information to build investor sentiment is particularly important. This paper chooses the sentiment of investors reflected by different search keywords and puts it into the GARCH model to predict the stock market volatility. This extended model, which contains the Baidu Index, is used to compare with the accuracy of common GARCH models. Similar to the existing research on Google Trends search data and forecast accuracy, the extended model with Baidu Index significantly improves the accuracy of forecasting Shanghai A-shares, with a 3% decline in MSE and a 15% decrease in U1 Theil. Shenzhen Stock Index, CSI 300 Index and CSI 100 Index also show that the conclusion has strong robustness. Therefore, it is clear that investors can use search engines such as Baidu to gather information to serve investment decision-making. Baidu Index can not only be used as an effective tool to reflect investor sentiment, but also help to roughly infer future market trends. In particular, during periods of high volatility, volatility in the return on stock markets will also intensify as investor searches increase. All in all, the Baidu Index reflects market volatility in almost real time. Especially in the period of economic fluctuation, it is of positive significance to use Baidu Index search data to identify potential risks and improve the forecast of stock market returns.

Reference

Ackert L F, Jiang L, Lee H S, et al. Influential investors in online stock forums. International Review of Financial Analysis. 2016; 45: 39–46.

Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I. Web search queries can predict stock market volumes. Plos One. 2012; 7(7):e40014.

Bordino I, Kourtellis N, Laptev N, Billawala Y. Stock trade volume prediction with Yahoo Finance user browsing behavior. In: Proc. 30th IEEE Intl. Conf. on Data Engineering (ICDE); 2014. p. 1168–1173

Caporin, M.; Poli, F.Building News Measures from Textual Data and an Application to Volatility Forecasting. Econometrics 2017, 5, 35.

Curme C, Preis T, Stanley HE, Moat HS. Quantifying the semantics of search behavior before stock market moves. Proc National Academy of Sciences. 2014; 111:11600–11605.

Da Z, Engelberg J, Gao P. In search of attention. The Journal of Finance. 2011; 66(5):1461–1499.

Dimpfl T, Jank S. Can internet search queries help to predict stock market volatility? European Financial Management. 2016; 22: 171-192.

Karim Rochdi, Marian Dietzel, (2015) "Outperforming the benchmark: online information demand and REIT market performance", Journal of Property Investment & Finance, Vol. 33 Issue: 2, pp.169-195

Kristoufek L. Can Google Trends search queries contribute to risk diversification? Sci Rep. 2013; 3.
Mao, Huina & Counts, Scott & Bollen, Johan. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data.
Preis T, Reith D, Stanley HE. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2010; 368(1933):5707-5719.
Pyo, Dong-Jin, Can Big Data Help Predict Financial Market Dynamics?: Evidence from the Korean Stock Market (June 30, 2017). East Asian Economic Review, Vol. 21, No. 2, pp. 147-165
R. H. Heiberger, Collective attention and stock prices: Evidence from google trends data on standard and poor's 100, PloS one 10 (8) (2015) e0135311.
Ramos, Henrique Pinto; Ribeiro, Kadja Katherine Mendes and Perlin, Marcelo Scherer. The forecasting power of internet search queries in the brazilian financial market. RAM, Rev. Adm. Mackenzie [online]. 2017, vol.18, n.2, pp.184-210.
Reference. Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. Financial Market and Portfolio Management, 25, 239-264.
Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports, 2013;Vol. 3, pp. 1684.
Usman, B., & Tandelilin, E. (2014). Internet Search Traffic and Its Influence on Liquidity and Returns of Indonesia Stocks: An Empirical Study. Journal of Indonesian Economy and Business. Vol. 29(3). Pp. 203-221.
Vlastakis N, Markellos RN (2012) Information demand and stock market volatility. J Bank Financ 36:1808-21.
Yahui Zhang, Difang Wan, Leiming Fu, (2013) "Impact of media on stock returns: an in-depth empirical study in China", Chinese Management Studies, Vol. 7 Issue: 4, pp.586-603.
Zhi Da, Joseph Engelberg and Pengjie Gao. Review of Financial Studies, 2015, vol. 28, issue 1, 1-32.