Search of Attention in Financial Market

Chong, Terence Tai Leung and Li, Chen

The Chinese University of Hong Kong

1 January 2020

Search of Attention in Financial Market

Terence Tai-Leung CHONG, Chen LI

Department of Economics, The Chinese University of Hong Kong

January 2020

Abstract: This study employs correlation coefficients and the factor-augmented vector autoregressive (FAVAR) model to investigate the relationship between the stock market and investors' sentiment measured by big data. The investors' sentiment index is constructed from a pool of relative keyword series provided by the Baidu Index. We target two composite stock indices, namely the Hang Seng Index and the Shanghai Composite Index. We first compute the Pearson product-moment correlation coefficient to find the degree of correlation between keywords and composite stock price indices. Then, we apply the FAVAR model to obtain the impulse response of stock price to the investors' sentiment index. Finally, we examine the leading effects of keywords on stock prices using lagged correlation coefficients. We obtain two main findings. First, a strong correlation exists between investors' sentiment and composite stock price: Second, before and after the launch of the Shanghai-Hong

Kong Stock Connect, the keywords affecting the fluctuation of the Hang Seng Index are different.

1 Introduction

With the advancement of technology, the Internet has become a very important source of information for investors. Over the past decade, investors have increasingly relied on online media for information. They conveniently make trading decisions by analyzing data obtained from the Internet (Brynjolfsson et al., 2013). Researchers have explored the impact of media on the real economy and other aspects. Luong et al. (2019) assess the impact of media on trade. Liu et al. (2017) discuss the impact of the Bo Xilai political scandal on policy uncertainty. In addition to receiving information provided by the media, investors also seek out information for themselves. Internet search engines, such as Baidu, Weibo, and Google, have already developed specific statistical methods to report search frequency on a daily or weekly basis. For instance, Google Trends provides search index services around the world. However, the most popular search engine in China is Baidu, which amassed approximately 0.87 billion users in 2017, reported by Soho Finance. According to Alexa, a web analytics firm, Baidu is the largest search engine used by Chinese people and the fourth largest website in the world. Another crucial search index is from Weibo, the application software of Alibaba Group, the largest communication and entertainment platform in China. This website provides useful instruments for conducting economic research. In recent years, many researchers have focused on the importance of the Internet search index to the market. Da et al. (2011) discuss the effects

of search volume on the stock market. They develop a new method to measure investor attention using search data in Google, which can be used to estimate the performance of IPO stocks. Preis et al. (2013) analyze how trading behavior can be captured using Google Trends. Wu and Brynjolfsson (2015) examine how Google Trends can be used to forecast housing prices and sales. They find that predictions are accurate when search frequency is included in their model. They employ search frequency to successfully explain three features of IPO stocks. Vosen and Schmidt (2011) contrast survey-based indicators and Google Trends in terms of private consumption prediction. Choi and Varian (2012) supply instructions for predicting different industrial sales by employing Google Trends. Xu (2014) forecasts stock price using Google Trends and Yahoo Finance. Challet and Ayed (2014) discuss whether Google Trends is superior to past price returns regarding the predictability of future price returns. Liu et al. (2011) adopt the lead-lag relationship approach to synthesize search indices and conduct the Granger causality test. Their results indicate that the data of search indices have remarkable predictive effectiveness for the annualized yield of the Shanghai Composite Index. For the analysis in the second step, they only use keywords which have leading correlation with stock prices. In this study, we filter keywords using normal and lead-lag relationship methods. Differentiating between the early and the late actions of two time series in the first two parts of discussion concerning the effects on the two stock price indices is

unnecessary, as the factor-augmented VAR (FAVAR) model adopted assumes the variables to be endogenous. This study divides the keyword series into two groups according to the strength of correlation between the keywords and composite stock price. If the Pearson coefficient between the keyword series and composite stock price exceeds 0.4, we consider the keywords to have a strong correlation with composite stock price; otherwise the keywords is considered to show a weak correlation. Furthermore, this study performs principle component analysis (PCA) instead of opting for a correlation synthesis method to extract factors from the keyword series to show investors' sentiment. We download data from the Shanghai Composite Index and Hang Seng Index through Yahoo Finance.

2 Data

Most literature constructs search indices by employing big data from Google Trends. Considering that Google terminated its search engine services in China in 2014, Google Trends may not be the best tool for obtaining the most comprehensive information about Chinese investors' search attention. This study relies on search frequency data from Baidu Index, as most Chinese investors are likely to search related keywords on Baidu. Identifying basic keywords relevant to the stock market is the first task in data collection. Liu et al. (2011) classify the keywords into three types: (1) investors' action index, which represents the keywords people search for if

they intend to open a stock market account; (2) investors' market quotation index, the keywords investors pay attention to if they have already opened a stock market account and intend to trade in financial markets; and (3) the economic situation index, the keywords signifying the economic performance or the government policy and international situation that investors prioritize. Following the study by Liu et al. (2011), our research also classifies keywords into different types to facilitate analysis, namely, investors' sentiment in the preliminary stage, which encompasses the keywords investors search for if they are about to enter the stock market; investors' sentiment on trading strategy, which indicates the actions investors take if they prepare to buy and sell stocks in the financial markets; investors' attention to the economy and policy, which denotes the economic situation and the government decisions that investors focus on; and investors' sentiment on other factors, such as company actions and strategies. After classifying the keywords, we select certain related keywords from financial news, reports, and applications on the basis of these four types as basic keywords. In general, these basic keywords often appear in financial media. Submitting basic keywords to the Baidu Index can generate several related words from one basic keyword, thus providing other keywords people also search for after they browse the keywords we inputted. Making use of the recommendation system of the website, we can accumulate a collection of related keywords from a relatively small set of basic keywords. After

applying this method and inputting the keywords successively, we have collected 331 keywords relevant to financial markets, including the name of the market index and the viewpoint of market participants. Figure 1 shows the search frequency of the term “stock market trend” (Chinese version). The resulting curve appears to resemble the trend observed in the Shanghai Composite Index. We collect the daily time series of the 331 keywords starting as early as 2011 from Baidu and retrieve data of the Shanghai Composite Index as well as the Hang Seng Index from Yahoo Finance. Table 1 presents part of the extracted keywords from the Baidu Index. The keywords gathered are in Chinese. For the purpose of clarity, we have translated them into English.

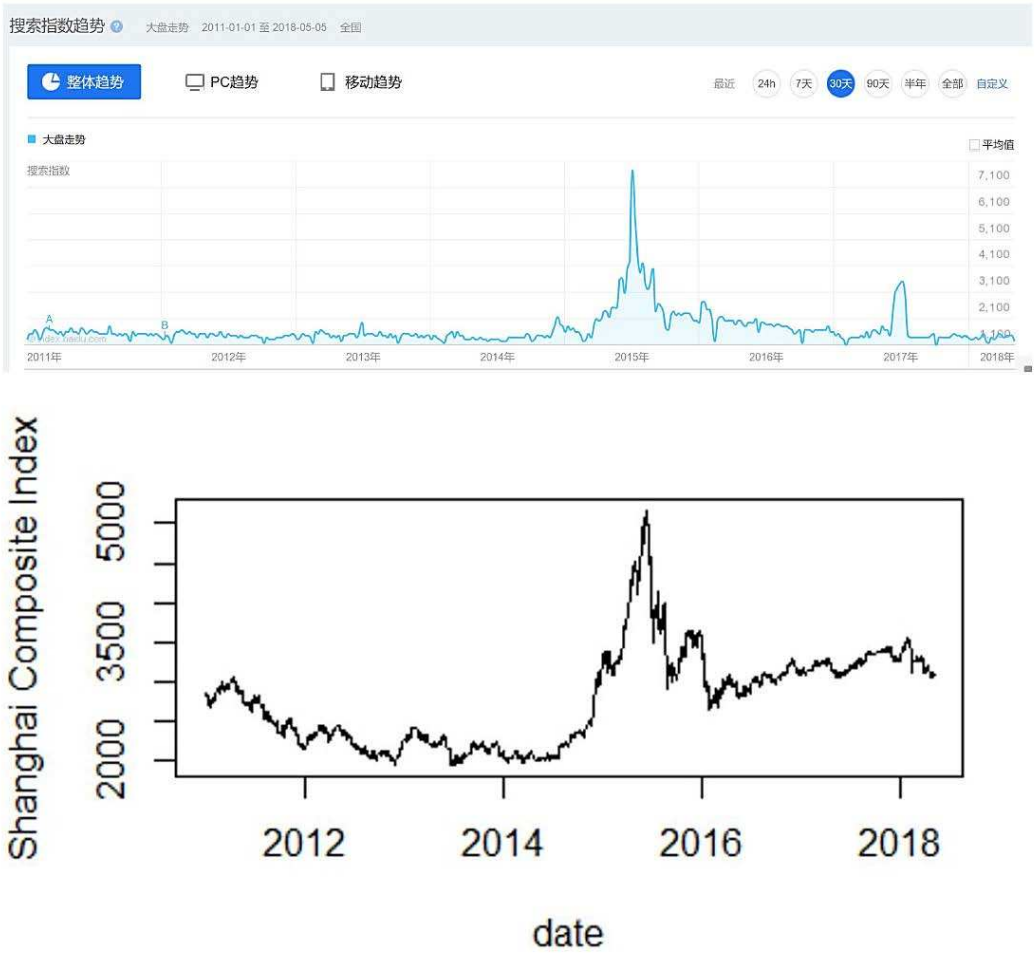


Figure 1: Single keyword index and the Shanghai Composite Index

	Keywords translated into English	Chinese version (original version)
1	Chinese news	中国新闻
2	Stock	股票
3	Dividend	红利
4	Stock introduction	股市入门
5	Prediction	预测
6	Caijing.com	财经网
7	Stock configuration	股票配置
8	Rise	涨停
9	Open security account	证券开户
10	Buy stock	买股票
11	Wealth	财富
12	Financial management	理财
13	Asset	资本
14	Security	金融证券
15	Demand	需求
16	Investment	投资
17	Exchange rate	人民币汇率
18	Financial market	金融市场
19	Asset management	资产管理
20	How to open an account	如何开户
21	CPI	CPI
22	Stock code	股票代码
23	Stock index	股指
24	K lines	K 线
25	Stock	Stock
26	Equity trading	融资
27	Brokerage	佣金
28	Crude oil	原油
29	Economy	经济
30	Bull market	牛市
31	Buying long	多头
...	...	...
331	Shanghai-Hong Kong Stock Connect	沪港通

Table 1: Part of the selected keywords (with English translation)

3 Methodology

The FAVAR model, proposed by Bernanke et al. (2005), is employed in this study. This model allows the extraction of relatively unobservable variables from a large pool of keyword data series and is proposed to include unobserved factors in autoregressive analysis. Compared with conventional vector autoregressive (VAR) models, the FAVAR model has richer information and can better explain certain macroeconomic phenomena. The FAVAR model is set out as follows:

$$\begin{bmatrix} y_t \\ F_t \end{bmatrix} = \Phi(L) \begin{bmatrix} y_{t-1} \\ F_{t-1} \end{bmatrix} + v_t, \quad \text{.....(1)}$$

$$X_t = \Lambda^f F_t + \Lambda^y y_t + e_t, \quad \text{.....(2)}$$

where y_t are variables that can be directly observed. F_t are unobserved variables that must be extracted from the related series. Here, F_t include variables about investors' sentiment in the preliminary stage and the trading stage, together with their sentiment on the economy and policy and on other factors, such as the financial situation. X_t is a vector comprising a pool of keywords and real economic data sets, which can be obtained from the Baidu Index and CEIC Data. Matrices Λ^f and Λ^y are factor loadings of dimensions conformable to X_t, $y_{t,}$ and F_t, whereas e_t are error terms assumed to have zero means.

Several studies have discussed macroeconomic problems by adopting

the FAVAR model. Some researchers examine the impact of monetary policy on the economy using macroeconomic variables with FAVAR models (He et al., 2013; Fernald et al., 2014). We utilize a FAVAR model to investigate the effects of the search index.

Principal Component Analysis (PCA)

PCA is a transformation method for converting a set of related data into a series of linearly uncorrelated factors. PCA is generally used in prediction and exploratory data analysis models. Our work deals with a huge number of interrelated keywords, and incorporating all keyword series in our models is impractical. PCA can be employed to extract the main features of all keyword series, which include the definitions of keywords, to construct the model. This study selects the number of factors on the basis of the results of variance explained, as illustrated in Figures 2, 3, and 5.

Lead-Lag relationship method

Liu et al. (2011) discuss the lead-lag relationship method in their study. They obtain factors by synthesizing keywords instead of using PCA. They first calculate the lagged correlation coefficient as the weight to synthesize the composite factor employed in their model:

$$\gamma_l^i = \frac{\Sigma_{t=1}^{n}(x_{t-l}^i - \bar{x}^i)(y_t - \bar{y})}{\sqrt{\Sigma_{t=1}^{n}\left(x_{t-l}^i - \overline{x^i}\right)^2 (y_t - \bar{y})^2}}, \ldots\ldots\ldots\ldots\ldots(3)$$

$$l = 0, \pm 1, \pm 2 \ldots \pm 10,$$

where y_t is the selected composite stock price . γ_l^i represents the lagged correlation coefficient of keyword i at a specific lag l. l^* is defined as the value of l of the keyword when the absolute value of γ_l^i is the largest, which means keyword i has the strongest correlation with y_t when its lag equals l. Then, they divide the keywords into three groups based on the feature of l^*. If l^* is positive, the keyword has a leading relationship with the composite stock price. If l^* is negative, the keyword and the composite stock price have a lagging relationship. If l^* equals zero, the keyword does not have a clear leading/lagging relationship with the composite stock price. Liu et al. (2011) select the keywords that have leading correlation with the composite stock price to synthesize the composite factor for their research. Considering that the FAVAR model allows for autocorrelation, our study utilizes all strongly correlated keywords instead of those pertaining to the first two parts for discussing the Shanghai Composite Index and the Hang Seng Index. In the last part of the study, where we forecast the stock price of a specific company, we select the leading correlated keywords in order to achieve better prediction performance, as seen in Liu et al. (2011).

We assume that composite stock price is driven by investors' sentiment

and financial/economic variables, such as exchange rates, short-term interest rates, and gold price. Investors' sentiments as latent factors must be distilled from the large pool of keyword series of the Baidu Index. This study also extracts common factors from a collection of real financial variables instead of feeding original data into the model. We have gathered 331 keywords as aforementioned. We do not analyze the effects of individual keywords on the stock market. Instead, we use a synthetic index representing investors' attention and sentiment, which can be obtained from the aggregate keyword time series. Here, we assume that the sentiment factor is a latent variable, which cannot be directly observed but can be inferred from a conglomeration of keywords. This study performs PCA to extract factors from keyword series. Liu et al. (2011) examine the lagging relationship between stock returns and the rate of change of keywords. They subsequently use keywords with a leading effect on stock return to create the index they need. In the section analyzing the keywords' effects on stock indices, we calculate the correlation (without lags) between the composite stock price index and the single keyword index, assuming that all variables are endogenous. Subsequently, we divide the keyword series into two parts: keywords having a strong correlation with composite stock price and those that are weakly correlated to composite stock price. If the Pearson coefficient between the keyword series and composite stock price is larger than 0.4, the keywords are considered to have a strong correlation with composite stock price;

otherwise, the keywords have a weak correlation. We discard the keywords with Pearson correlation coefficients of less than 0.4, thus retaining keyword series with high correlation with stock price. Unlike the research by Liu et al. (2011), which uses the synthesis method, our study employs PCA to extract factors. We compile four indices based on the definitions of the remaining keywords before performing PCA to improve the extraction and explanation of factors: (1) investors' sentiment in the preliminary stage, including how to open a securities account and how to speculate in shares; (2) investors' sentiment in the trading stage, including market quotations, dividends, and the amount of increase; (3) investors' attention to policy and the economy, including financial news and international situation; and (4) investors' attention to other factors, including keywords about companies' actions and performance in financial markets. These four indices are not independent from each other, as investors may search for all of them to formulate an improved trading strategy.

	Keywords (original version)	Keywords (translation in English)
1	行情	market
2	收益	earnings
3	A股	A-share
4	银行股	bank shares
5	新浪财经	Sina Finance
6	空头	short position
7	多头	long position
8	金融市场	financial market
9	大盘走势	market trends
10	黑马	dark horse
...	...	...

Table 2: Keywords for investors' sentiment index in the trading stage

	Keywords (original version)	Keywords (translation in English)
1	股市入门	introduction to the stock market
2	证券开户	open stock account
3	K线图怎么看	how to analyze K lines
4	买股票	buy stock
5	炒股经验	knowledge on financial market
6	如何开户	how to open a financial account
7	股票查询	check stock value
8	股票代码查询	check stock code
9	股票推荐	stock recommendations
10	股票开户	open stock account ... ...
...		

Table 3: Keywords for investors' sentiment index in the preliminary stage

	Keywords (original version)	Keywords (translation in English)
1	中国新闻	Chinese news
2	中国经济	Chinese economy
3	国际形势	international situation
4	人民币汇率	exchange rate of RMB
5	经济数据	economic data
6	发展	development
7	税收	tax revenue
8	通货膨胀率	inflation rate
9	CPI	CPI
10	原油	crude oil
...	...	...

Table 4: Keywords for investors' sentiment index on the economy and policy

	Keywords (original version)	Keywords (translation in English)
1	银行贷款	bank loan
2	风险管理	risk management
3	信息披露	information disclosure
4	风险评估	risk evaluation
5	借壳上市	back-door listing
6	信用卡	credit card
7	风险敞口	risk exposure
8	不良贷款	bad debts
9	基金公司	fund company
10	重组	acquisition and reorganization
...	...	...

Table 5: Keywords for investors' sentiment index on other factors

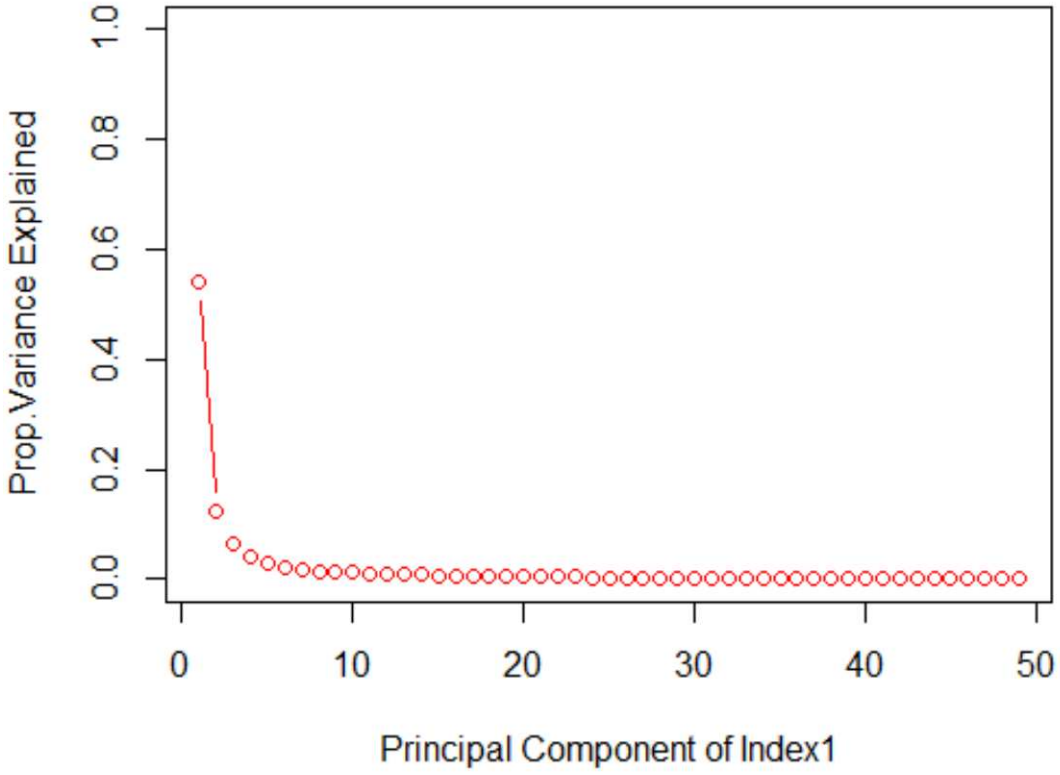
	$\rho > 0.4$	$\rho > 0.5$	$\rho > 0.6$	$\rho > 0.7$
	中国新闻(Chinese news)	中国新闻(Chinese news)	平台(channel)	平台(channel)
	股票(stock)	股票(stock)	易方达(efunds)	东方财富网(eastmoney.com)
	平台(channel)	平台(channel)	和讯网(hexun.com)	新浪财经(finance.sina.com)
	股市入门(introduction to the stock market)	涨停(limit up)	行情(price/performance)	股票代码(stock code)
	财经网(caijing.com)	易方达(efunds)	东方财富网(eastmoney.com)	同花顺(10jqka.com.cn)
	涨停(limit up)	证券开户(open stock account)	银行贷款(bank loan)	k线图(K line)
	易方达(efunds)	解禁股(restricted stock circulation)	炒股软件(stock software)	雪球(xueqiu.com)
	证券开户(open stock account)	炒股(trading in stock market)	h股(h share)	股票市场(stock market)
	解禁股(restricted stock circulation)	k线图怎么看(how to analyze K line)	新浪财经(finance.sina.com)	股票代码查询(stock code inquiry)
	财新网(caixin.com)	买股票(buy stock)	基金(fund)	stock
	炒股	和讯网(hexun.com)	资本市场(capital market)	股票交易(stock trading)
	k线图怎么看(how to analyze K line)	行情(price/performance)	大盘行情(stock market performance)	股权(equity)
	嘉实(Harvest Fund)	东方财富网(eastmoney.com)	如何开户	成交量(trading volume)
	买股票(buy stock)	投资(investment)	股票新闻(stock news)	股份(share)
	和讯网(hexun.com)	收益(profit)	股票代码(stock code)	私募(private placement)
	行情(price)	中国证券网(cnstock.com)	指数(index)	市盈率(P/E ratio)
	资本(capital)	银行贷款(bank loan)	同花顺(10jqka.com.cn)	复盘(compound)
	东方财富网(eastmoney.com)	净值(net value)	k线图	k线(K line)
	投资(investment)	炒股软件(stock software)	股票行情(stock market price)	什么是k线(What is the K line)
	收益(profit)	银行股(stock of banking)	雪球(xueqiu.com)	手机炒股软件哪个好(best mobile phone stock software)
	中国证券网(cnstock.com)	h股(h share)	股票市场(stock market)	换手率(hand turnover rate)
	银行贷款(bank loan)	新浪财经(finance.sina.com)	财经新闻(economic news)	港股通交易规则(Hong Kong Stock Connect Trading Rules)
	a股	风险控制(risk management)	股票代码查询(stock code inquiry)	港股交易时间(Hong Kong stock trading hours)
	...	...	...	...
Total	180	118	67	23

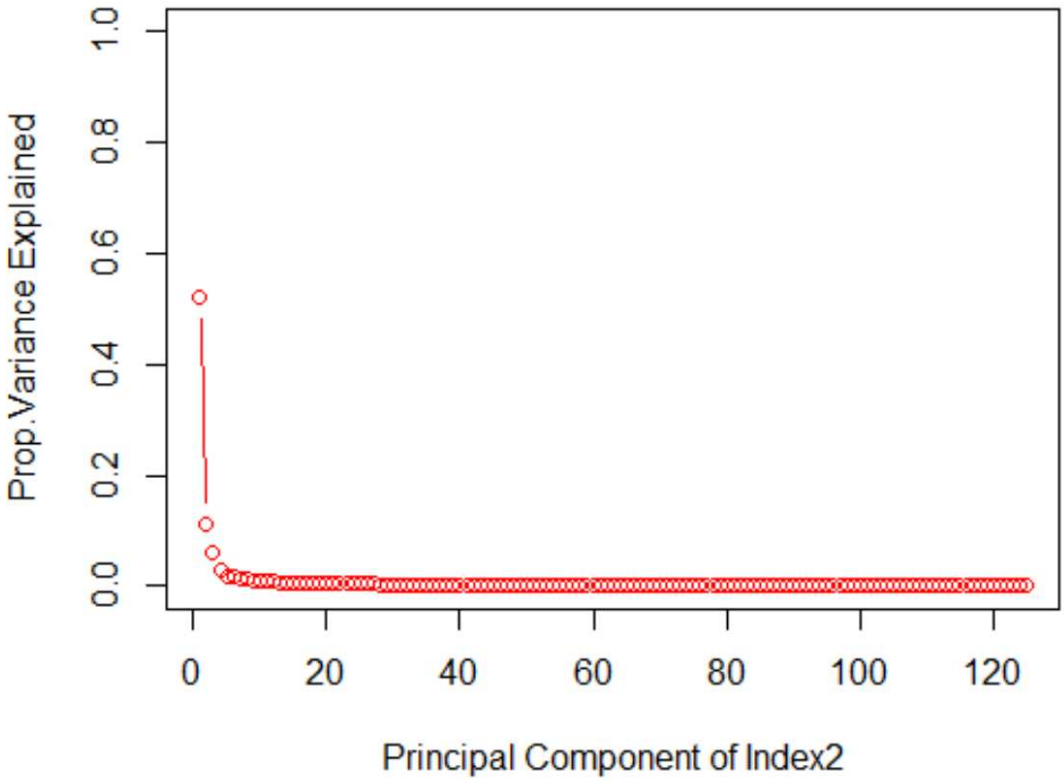
Table 6: Keywords that have a strong correlation with the Shanghai Composite Index

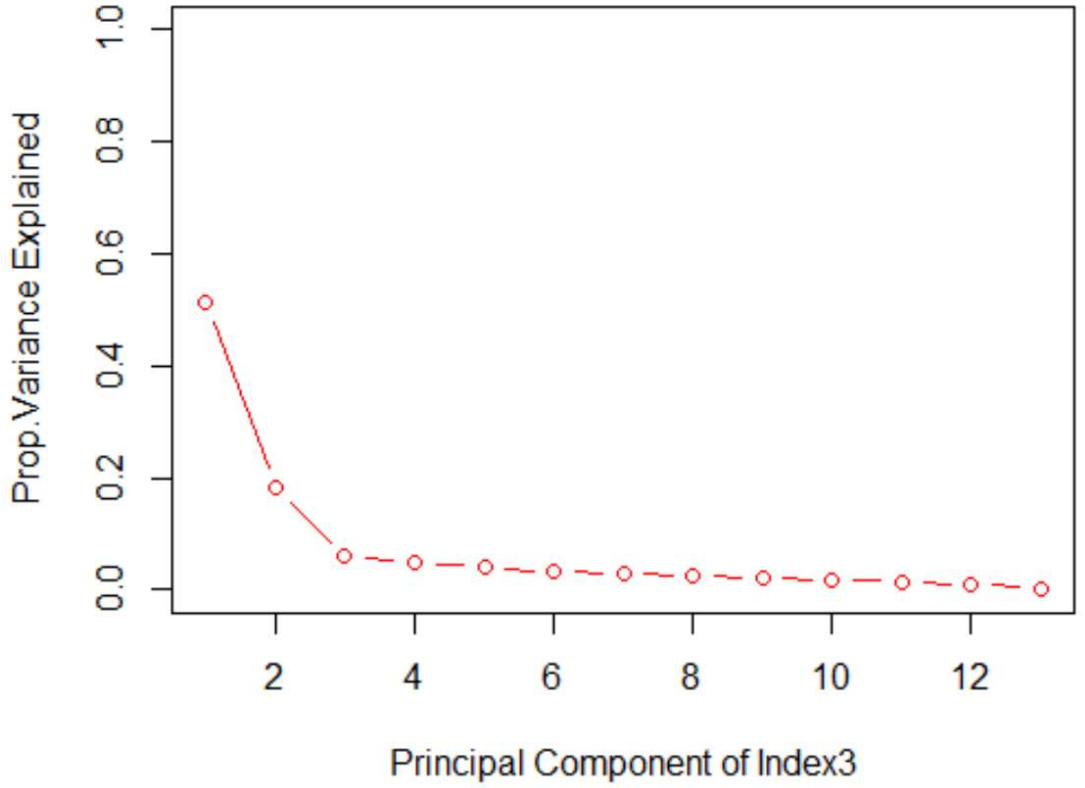
Investors' sentiment in trading	Investors' sentiment in preliminary stage	Investors' sentiment on economy and policy	Investors' sentiment on other factors
股票(stock)	股票(stock)	中国新闻(Chinese News)	资本(capital)
平台(channel)	财经网(caijing.com)	通货膨胀率(inflation rate)	银行贷款(bank loan)
股市入门(introduction to the stock market)	涨停(limit up)	汇率(exchange rate)	风险控制(risk management)
证券开户(open stock account)	易方达(efunds)	原油(crude oil)	信息披露(information disclosure)
炒股(trade in stock market)	解禁股(restricted stock circulation)	能源(energy)	风险评估(risk assessment)
k线图怎么看(how to analyze K line)	财新网(caixin.com)	去杠杆(de-leveraging)	基金(fund)
买股票(buy stock)	炒股(trade in stock market)	美元(dollar)	基金公司(fund company)
a股(a share)	嘉实(Harvest Fund)	保监会(China Insurance Regulatory Commission)	期权(option)
炒股软件(stock software)	买股票(buy stock)	银监会(Banking Regulatory Commission)	上市公司信息披露办法(listed company information disclosure method)
大智慧(gw.com.cn/)	和讯网(hexun.com)	央行(Central bank)	资产管理(asset management)
如何开户(how to open a stock account) account	行情(price/performance)	宏观经济(macro-economy)	互换(swap)
股票代码(Stock code)	东方财富网(eastmoney.com)	恒生指数(Hang Seng Index)	外资(foreign investment)
...	...	...	...

Table 7: Four types of keywords that have a strong correlation with the composite index

The results of explained variance from the PCA are reported in Figure 2. We extract five factors in total from the four keyword series based on the results in Figure 2, explaining 60%–80% of the information. We also extract two factors from real economic and financial series as control variables by employing PCA. In addition, we calculate the Pearson correlation coefficient between factors and stock price, and the results indicate a high correlation ranging from 0.6 to 0.8.


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Principal Component of Index1


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Principal Component of Index2



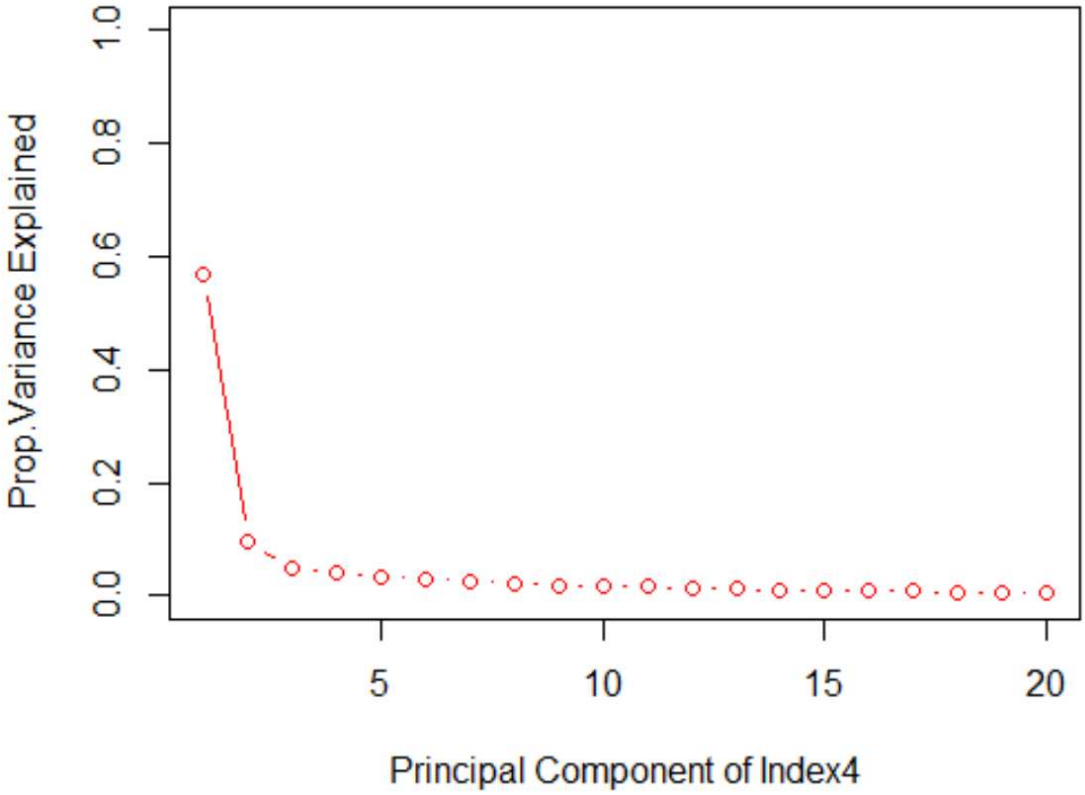


Figure 2: Explained variance of PCA

Before estimating using the FAVAR model, we transform our data as follows:

$$y_t = 100 \times \ln\left(\frac{P_t}{P_{t-1}}\right), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

$$f_t = I_t - I_{t-1}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

where P_t is the log-differenced composite price, and I_t is the index we extracted from the keyword series through PCA. We transform factors by first-order differencing instead of log-differencing because of the negative value in I_t after conducting PCA. The FAVAR model is as follows:

$$\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \Phi(L)\begin{bmatrix} Y_{t-1} \\ F_{t-1} \end{bmatrix} + v_{t,} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$X_t = \Lambda^f F_t + \Lambda^y Y_t + e_{t,} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

where Y_t contains the rate of change of the Shanghai Composite Index. F_t includes the change of investors' sentiment factors and other control factors after first-order differencing. The assumption is that composite stock price and search frequency are endogenous because they can influence each other under a shock to any one of them. X_t is a vector comprising a collection of keywords and real economic data sets, which can be obtained from the Baidu Index and CEIC Data. Matrices Λ^f and Λ^y are the factor loadings of dimensions conformable to X_t, Y_t, and F_t, whereas e_t are error terms

assumed to have zero means.

4 Results

Tables 10–12 present parts of the results of the FAVAR model. Table 10 shows the estimation performance when investors' sentiment in the preliminary stage is the dependent variable. Here, X1 denotes the factor of investors' sentiment in the preliminary stage. *Return* represents the rate of change of the Shanghai Composite Index. The return of stock price has a noteworthy effect on investors' decisions of whether to open a stock market account. The estimated coefficient of *return* is positive. When the stock market is booming, many investors become interested in opening accounts in the stock market. Consequently, investors' search frequency for keywords related to opening stock accounts increases. Table 11 presents the regression results where investors' sentiment in the trading stage is the dependent variable and can be influenced by stock price, the economic situation, and companies' actions and performance. Table 11 shows the significant effects of *return*, X4, X5, and X7 on X2, where X2 represents investors' sentiment in the trading stage. X3 and X4 signify investors' attention on the economy and government policy, respectively. X5 represents investors' attention to other factors, including company actions and performance. X6 and X7 are the factors extracted from control variables describing the financial and economic situations. We cannot simply interpret that a positive shock to

investors' sentiment in the trading stage can cause a positive change in return. Negative/positive news not only increases search frequency but also exerts different impacts on stock return. We do not separate the keywords according to negative or positive effects. Whether the correlation between search frequency and stock return is positive or negative depends on the period and the keywords selected, thus warranting further discussion in the future. Table 12 presents the results where *return* becomes the dependent variable. The effects of the variables on stock price returns are less significant than the other two regressions. One explanation is that, given the daily data, a shock to investors' search frequency in one day is inconsequential in influencing composite stock price in the broader financial market. For example, the increase in X1 indicates that many investors plan to enter the financial market, but the actual increase in stock market accounts is probably more trifling than that of X1. The value added to the stock market by this factor is also small. Moreover, we find that investors' sentiment in the trading stage has a significant impact on the return of the stock market. This finding is in line with our expectations. In addition, the performance of X4 reveals that investors pay much attention to the economic situation and policy on implementing trading strategies in the stock market.

	AIC(n)	HQ(n)	SC(n)	FPE(n)
1	10	5	2	10

Table 8: Selection method

	AIC(n)	HQ(n)	SC(n)	FPE(n)
1	-5.5568	-5.4749	-5.3351	0.0039
2	-5.8102	-5.6556	-5.3915	0.0030
3	-5.9649	-5.7375	-5.3492	0.0026
4	-6.0910	-5.7908	-5.2783	0.0023
5	-6.1959	-5.8230	-5.1862	0.0020
6	-6.2162	-5.7705	-5.0095	0.0020
7	-6.2215	-5.7031	-4.8178	0.0020
8	-6.2329	-5.6417	-4.6322	0.0020
9	-6.2415	-5.5775	-4.4438	0.0019
10	-6.2629	-5.5262	-4.2682	0.0019

Table 9: Information criteria

	Estimate	Std. Error	t value	Pr(>|t|)
return.l1	0.0630	0.0161	3.9095	0.0001***
X1.l1	-0.1167	0.0361	-3.2357	0.0012**
X2.l1	0.0376	0.0264	1.4244	0.1545
X3.l1	-0.0817	0.0399	-2.0498	0.0405*
X4.l1	-0.0725	0.0454	-1.5958	0.1107
X5.l1	-0.1270	0.0284	-4.4673	0.0000***
X6.l1	-0.0187	0.0379	-0.4924	0.6225
X7.l1	-0.0989	0.0338	-2.9254	0.0035**
return.l2	0.0134	0.0162	0.8276	0.4080
X1.l2	-0.1663	0.0364	-4.5699	0.0000***
X2.l2	0.0332	0.0262	1.2663	0.2056
X3.l2	-0.0794	0.0398	-1.9947	0.0462*
X4.l2	-0.0264	0.0449	-0.5867	0.5575
X5.l2	-0.0369	0.0285	-1.2954	0.1953
X6.l2	-0.0287	0.0380	-0.7565	0.4495
X7.l2	-0.0770	0.0338	-2.2805	0.0227*
const	0.0037	0.0201	0.1821	0.8555

Table 10: Regression: X1 → *return*

	Estimate	Std. Error	t value	Pr(>|t|)
return.l1	0.1000	0.0284	3.5155	0.0004***
X1.l1	-0.0206	0.0636	-0.3230	0.7467
X2.l1	-0.0867	0.0465	-1.8638	0.0625.
X3.l1	-0.1314	0.0703	-1.8685	0.0619.
X4.l1	-0.2152	0.0802	-2.6837	0.0073**
X5.l1	-0.2872	0.0502	-5.7260	0.0000***
X6.l1	-0.0402	0.0669	-0.6016	0.5475
X7.l1	-0.1923	0.0597	-3.2228	0.0013**
return.l2	0.0934	0.0285	3.2750	0.0011**
X1.l2	-0.0261	0.0642	-0.4069	0.6841
X2.l2	-0.0806	0.0463	-1.7429	0.0815.
X3.l2	-0.0975	0.0702	-1.3874	0.1655
X4.l2	-0.1952	0.0793	-2.4622	0.0139*
X5.l2	-0.1246	0.0502	-2.4810	0.0132*
X6.l2	-0.0703	0.0670	-1.0486	0.2945
X7.l2	-0.1431	0.0596	-2.4008	0.0165*
const	0.0068	0.0354	0.1908	0.8487

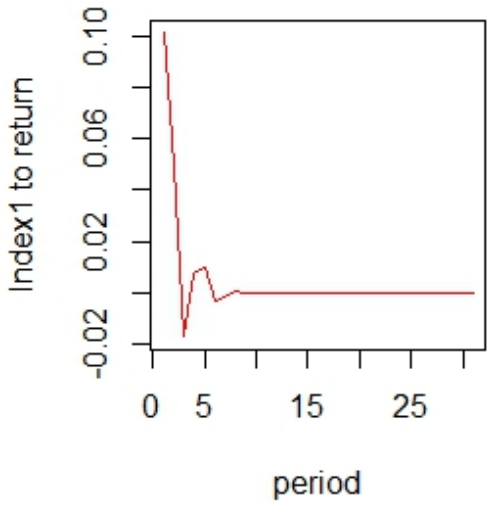
Table 11: Regression: X2 → *return*

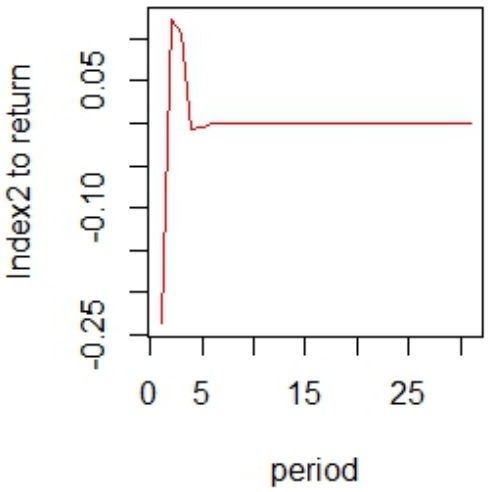
The impulse response results in Figure 3 reveal the relationship between stock returns and sentiment indices of sampled investors. Index1 and Index2 in Figure 3 represent investors' sentiment in the preliminary and the trading stages, respectively. Index3-1 and Index3-2 show the two factors extracted from investors' attention to the economy and policy. Finally, Index4 represents the factor extracted from investors' attention to other factors. Given a shock to a search index, the stock returns suddenly fluctuate in the following few days, and vice versa. As previously discussed, we cannot

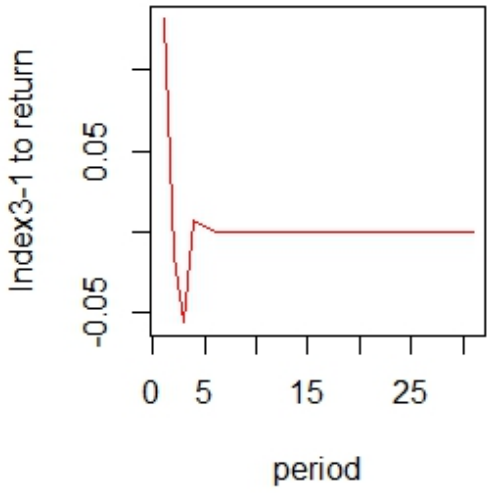
simply interpret that a positive shock to investors' sentiment can cause a positive/negative change in stock return. Both negative news and economic improvements can increase search frequency but have different impacts on stock returns.

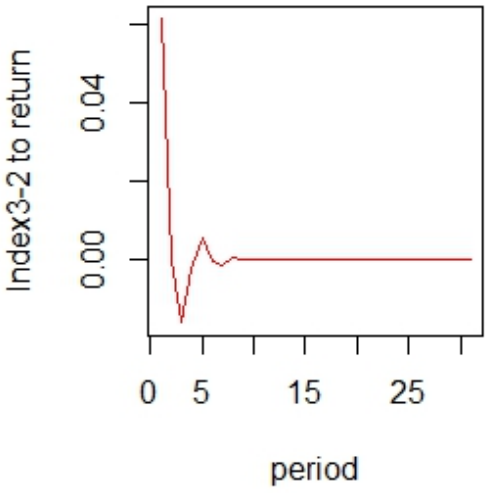
	Estimate	Std. Error	t value	Pr(>|t|)
return.l1	0.0497	0.0261	1.9002	0.0576
X1.l1	0.0436	0.0585	0.7450	0.4564
X2.l1	-0.0437	0.0428	-1.0204	0.3077
X3.l1	-0.0640	0.0647	-0.9901	0.3223
X4.l1	0.0029	0.0737	0.0395	0.9685
X5.l1	0.0387	0.0461	0.8390	0.4016
X6.l1	0.0087	0.0615	0.1411	0.8878
X7.l1	-0.0593	0.0549	-1.0803	0.2801
return.l2	-0.0236	0.0262	-0.8988	0.3689
X1.l2	-0.0471	0.0591	-0.7969	0.4256
X2.l2	0.1216	0.0425	2.8574	0.0043**
X3.l2	-0.0024	0.0646	-0.0371	0.9704
X4.l2	0.1565	0.0729	2.1464	0.0320*
X5.l2	-0.0706	0.0462	-1.5297	0.1263
X6.l2	0.0184	0.0616	0.2990	0.7650
X7.l2	-0.0646	0.0548	-1.1796	0.2383
const	0.0065	0.0326	0.1986	0.8426

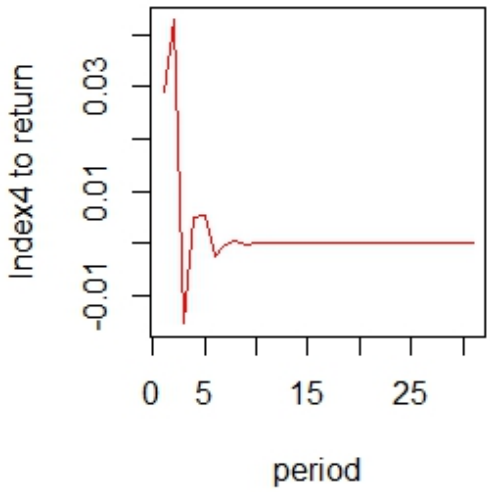
Table 12: Regression: Return → X1 + X2 +…

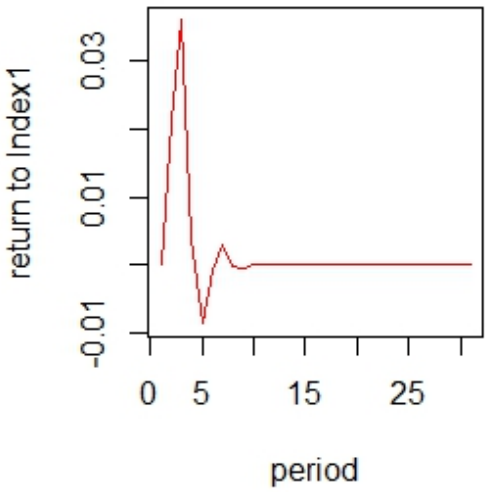

Index1 to return
-0.02
0.02
0.06
0.10
0
5
15
25
period

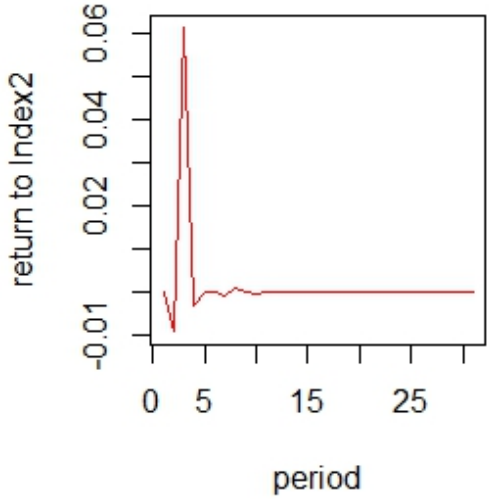

Index2 to return
-0.25
-0.10
0.05
0
5
15
25
period

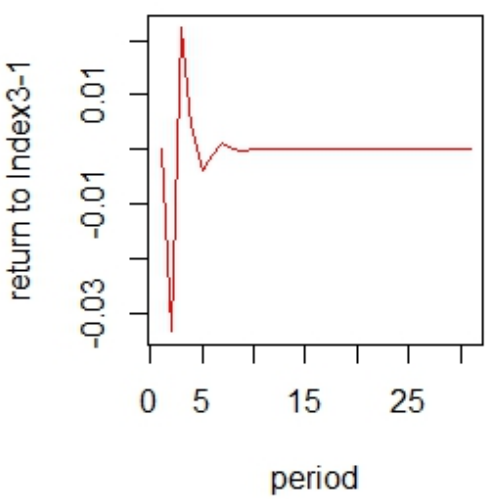

Index3-1 to return
-0.05
0.05
0
5
15
25
period


Index3-2 to return
0.00
0.04
0
5
15
25
period


Index4 to return
-0.01
0.01
0.03
0
5
15
25
period


return to Index1
-0.01
0.01
0.03
0
5
15
25
period


return to Index2
-0.01
0.02
0.04
0.06
0
5
15
25
period


return to Index3-1
-0.03
-0.01
0.01
0
5
15
25
period

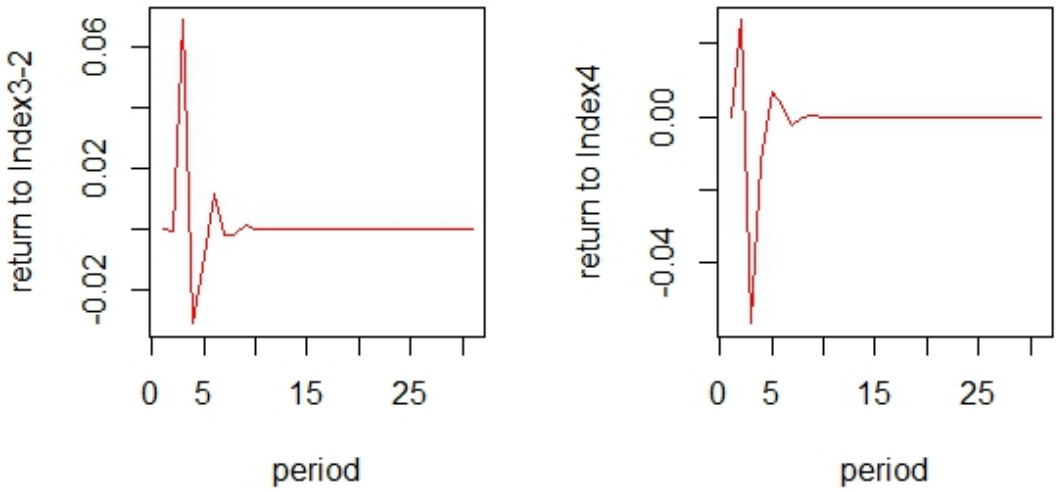


Figure 3: Impulse response of returns on the Shanghai Composite Index and factors

5 The relationship between the Hang Seng Index and Search Indices

Apart from the Shanghai Composite Index, this study also explores the effects of the same keyword series on the Hong Kong financial market. The Hang Seng Index represents the Hong Kong stock market in this research. As opposed to the discussion on the Shanghai Composite Index and search keywords, we divide the sample period into two. In April 2014, the Chinese government first announced the launch of the Shanghai-Hong Kong Stock Connect to promote trading between Shanghai and Hong Kong financial markets. We select the date of announcement, 9 April 2014, as the critical day separating the two periods. The first period is from 2011 to 9 April 2014,

and the second period is from 10 April 2014 to 2018. This study finds that the keyword series having a strong relationship with the Hang Seng Index can differ across the two periods. The main difference is evidently correlated with the Shanghai-Hong Kong Stock Connect. Additional keywords associated with the Shanghai-Hong Kong Stock Connect affect the changes in the Hang Seng Index after the announcement of the inception of the Shanghai-Hong Kong Stock Connect. Furthermore, we can observe the regularity of this phenomenon and propose an explanation for it. Tables 13–15 show the Pearson correlation coefficients between keywords and the Hang Seng Index in different periods. Three findings are obtained. First, the number of keywords having a strong correlation with the Hang Seng Index is smaller than that with the Shanghai Composite Index because our keyword series is manually selected, implying that there are some subjective elements in data selection. The keywords are recommended by the Baidu Index, the most popular search engine among Chinese users. Most keywords are selected on the basis of their association with Chinese companies. Although many Chinese companies participate in the Hong Kong financial market, Chinese investors tend not to trade the stocks of certain foreign companies due to relative unfamiliarity; such foreign companies may not have a discernible relationship with the selected keywords. Second, we observe different keyword types influencing the Hang Seng Index during different periods. Tables 13–15 distinguish between different types of keywords in the

two periods according to correlation coefficients. Investors' sentiment on trading and the economy are the general factors influencing the Hang Seng Index. Tables 13–15 indicate that keywords related to the Shanghai-Hong Kong Stock Connect are strongly correlated with the Hang Seng Index in the second period. Third, if we compare the related keyword coefficients for the Shanghai Composite Index and the Hang Seng Index, investors' investment in the A-share market is more tightly linked to the developments of Southbound trading (Hong Kong Stock Connect), whereas investment in the Hong Kong financial market is more closely associated with those of Northbound trading (Shanghai Stock Connect). The Southbound and Northbound trading discussed here are parts of the mechanism of the Shanghai-Hong Kong Stock Connect scheme.

$\rho > 0.4$	Before 20140409	After 20140409
	中国经济(Chinese economy)	中国经济(Chinese economy)
	国际形势(International situation)	经济(economy)
	美股(US stock market)	全球股市行情(global stock market situation)
	人民币汇率(RMB exchange rate)	全球股市(global stock market)
	全球股市指数(global stock market index)	原油(crude oil)
	全球股市行情(global stock market situation)	贵金属(precious metal)
	全球股市(global stock market)	外资(foreign investment)
	恒生指数(Hang Seng Index)	股票代码(stock code)
	道琼斯指数(Dow Jones Index)	股票市场(stock market)
	宏观经济(macro-economy)	零售(retail)
	恒生指数是什么(what is Hang Seng Index)	股票交易(stock trading)
	外汇(foreign exchange)	家电(household appliances)
	经济(economy)	k线(K line)
	家电(household appliances)	腾讯股票(Tencent stock)
	交易(trade)	牛市熊市(bull market, bear market)
	电力设备(electric equipment)	股票估值(Stock valuation)
	传媒行业(media industry)	股价(stock price)
	主线(main line)	财务报表(financial statements)
	操盘(trader)	沪港通资金流向(Shanghai-Hong Kong Stock Connect Capital Flow)
	黑马股(dark horse stock)	港股交易时间(Hong Kong stock trading hours)
	风险敞口(risk exposure)	港股行情(Hong Kong stock market situation)
	指数(index)	香港交易所(Hong Kong Stock Exchange)
	股指期货开户(open stock index futures account	沪股通(Shanghai Stock Connect)
	组合(combination)	沪股通资金流向(Shanghai Stock Connect Capital Flow)

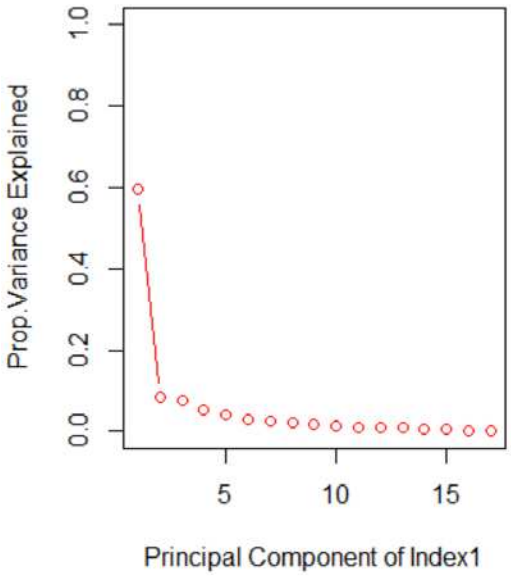
Table 13: Estimation of the relationship between the Hang Seng Index and the search indices using the Pearson correlation method

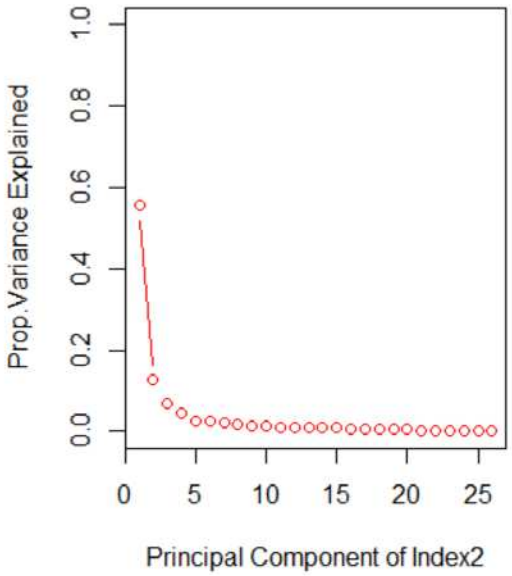
$\rho > 0.5$	Before 20140409	After 20140409
	中国经济(Chinese economy)	中国经济(Chinese economy)
	组合(combination)	股票代码(stock code)
	家电(household appliances)	股票市场(stock market)
	经济(economy)	经济(economy)
	美股(US stock market)	原油(crude oil)
	操盘(trader)	腾讯股票(Tencent stock)
	全球股市指数(global stock market index)	牛市熊市(bull market, bear market)
	全球股市行情(global stock market situation)	全球股市(global stock market)
	全球股市(global stock market)	股价(stock price)
	恒生指数(Hang Seng Index)	财务报表(financial statements)
	道琼斯指数(Dow Jones Index)	沪港通资金流向(Shanghai-Hong Kong Stock Connect Capital Flow)
		沪股通(Shanghai Stock Connect)
		沪股通资金流向(Shanghai Stock Connect Capital Flow)

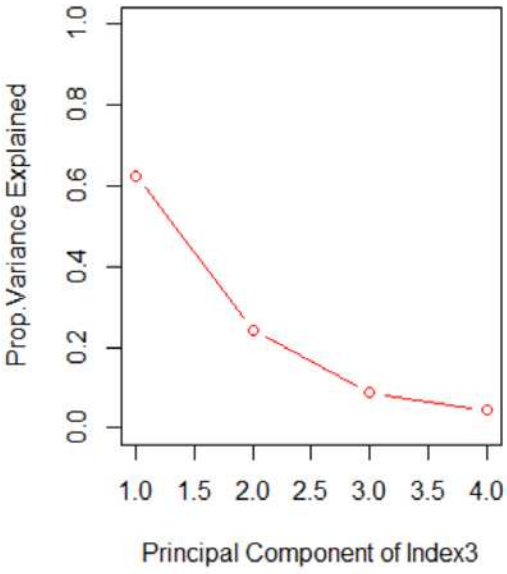
Table 14: Estimation of the relationship between the Hang Seng Index and the search indices using the Pearson correlation method

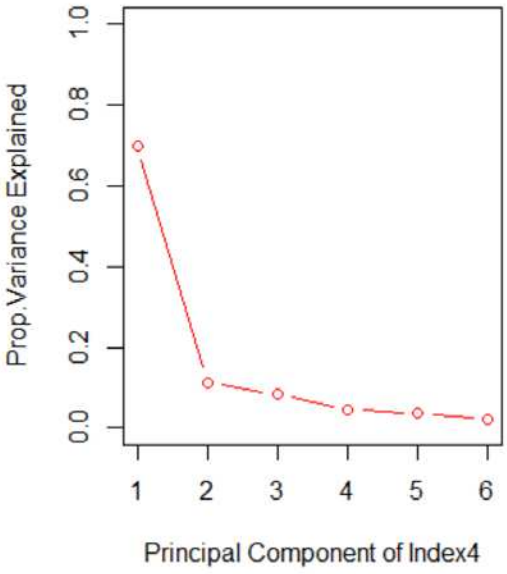
$\rho > 0.6$	Before 20140409	After 20140409
	美股(US stock market)	中国经济(Chinese economy)
	组合(combination)	股票市场(stock market)
	全球股市指数(global stock market index)	全球股市(global stock market)
	全球股市行情(global stock market situation)	股价(stock price)
	全球股市(global stock market)	沪港通资金流向(Shanghai-Hong Kong Stock Connect Capital Flow)

Table 15: Estimation of the relationship between the Hang Seng Index and the search indices using the Pearson correlation method


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
Principal Component of Index1


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Principal Component of Index2


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component of Index3


Prop.Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Principal Component of Index4

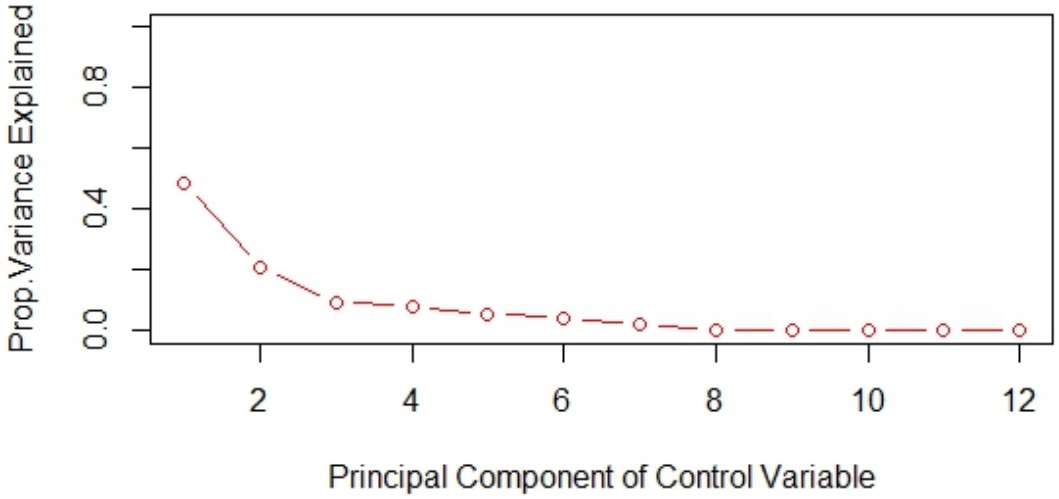
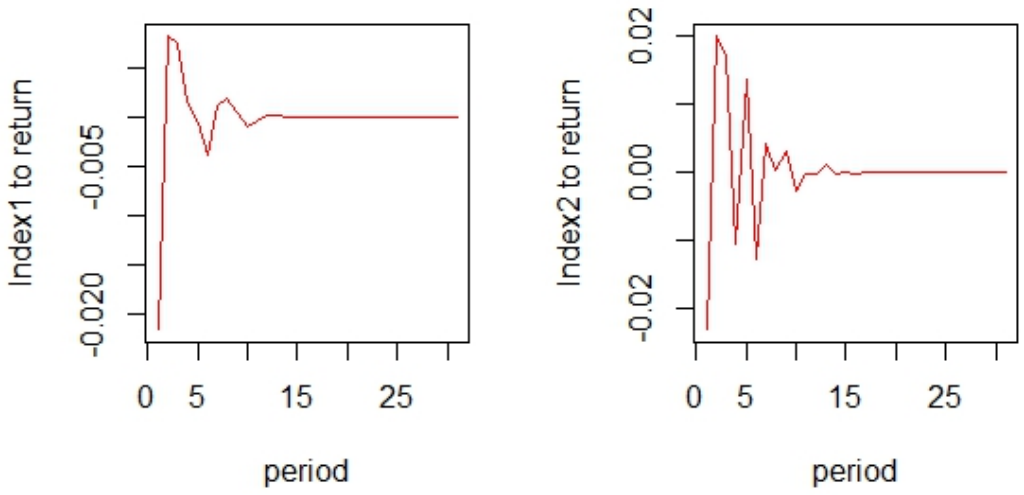
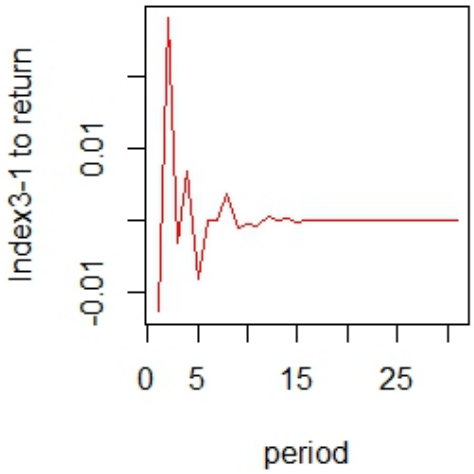
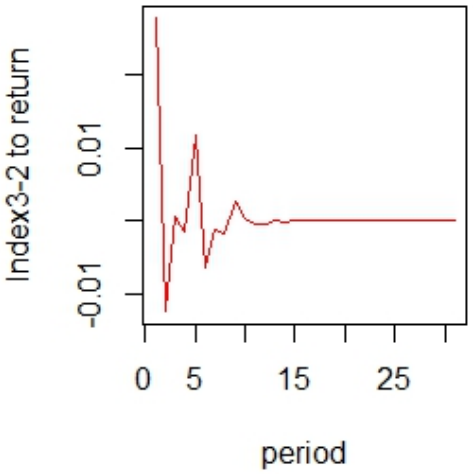
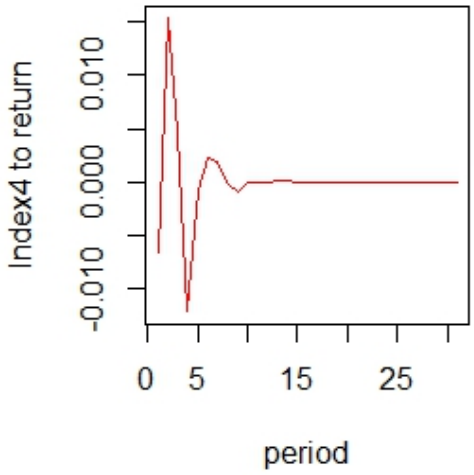


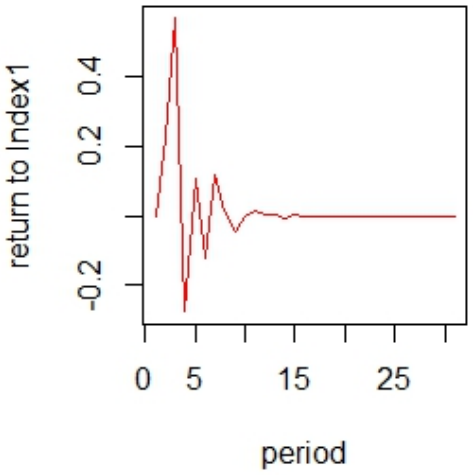
Figure 4: Explained variance by PCA




Index3-1 to return
-0.01
0.01
0
5
15
25
period


Index3-2 to return
-0.01
0.01
0
5
15
25
period


Index4 to return
-0.010
0.000
0.010
0
5
15
25
period


return to Index1
-0.2
0.2
0.4
0
5
15
25
period

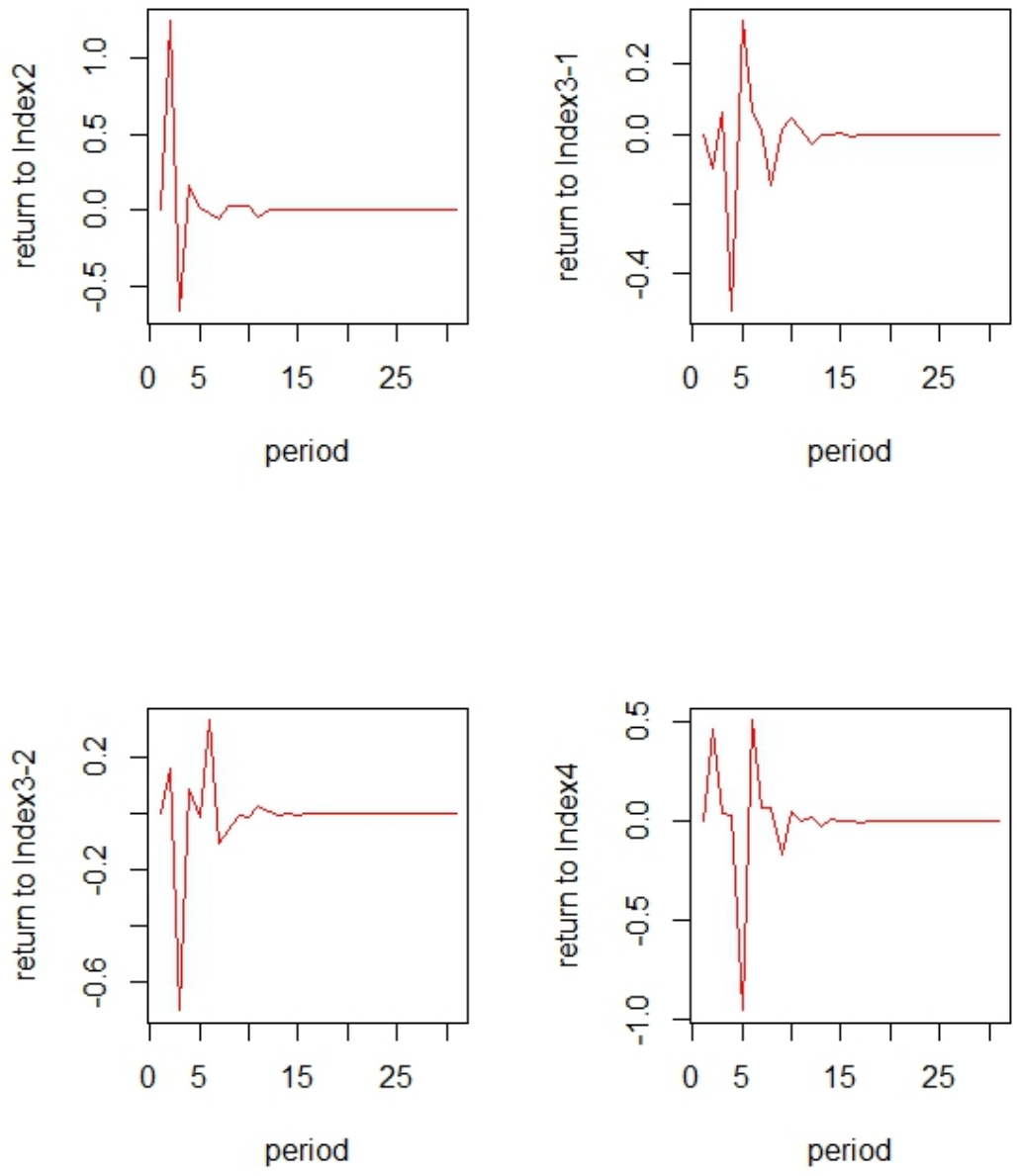


Figure 5: Impulse response results of returns on the Hang Seng Index and factors

Following the existing literature, this research also analyzes the impulse response of the Hang Seng Index to various factors. A noticeable relationship between investors' sentiment and the Hong Kong financial market is

observed; however, as above-mentioned, we cannot simply interpret that a positive shock to investors' sentiment can cause a positive or negative change in returns.

6 Further Discussion: Prediction of a Single Stock

The impact of the search indices on the stock market has been confirmed in previous sections. Drawing on the results obtained, this study constructs a predictive model on the firm level by incorporating the search indices. We select the stock of Tencent (0700.HK), a popular stock that attracts the attention of many investors, as the object of our research. In recent years, Tencent has seen a substantial increase in stock price due to the growth and development of the company. This study intends to predict Tencent's stock performance by developing a predictive model and using rolling forecasting methods. Several regression models are available for selection. Li and Yang (2013) predict investment risk and return by using a linear regression model. Here, we use FAVAR models to forecast stock price. We collect relative keyword series from Google Trends instead of the Baidu Index because the stock in question (0700.HK) is listed in Hong Kong and has an international mix of investors. Google Trends may be more representative and could include further information for explaining investors' sentiment. We select 52 keywords in total, as presented in Table 16. We also choose price-to-earnings (PE) ratios, the Hang Seng Index, earnings per

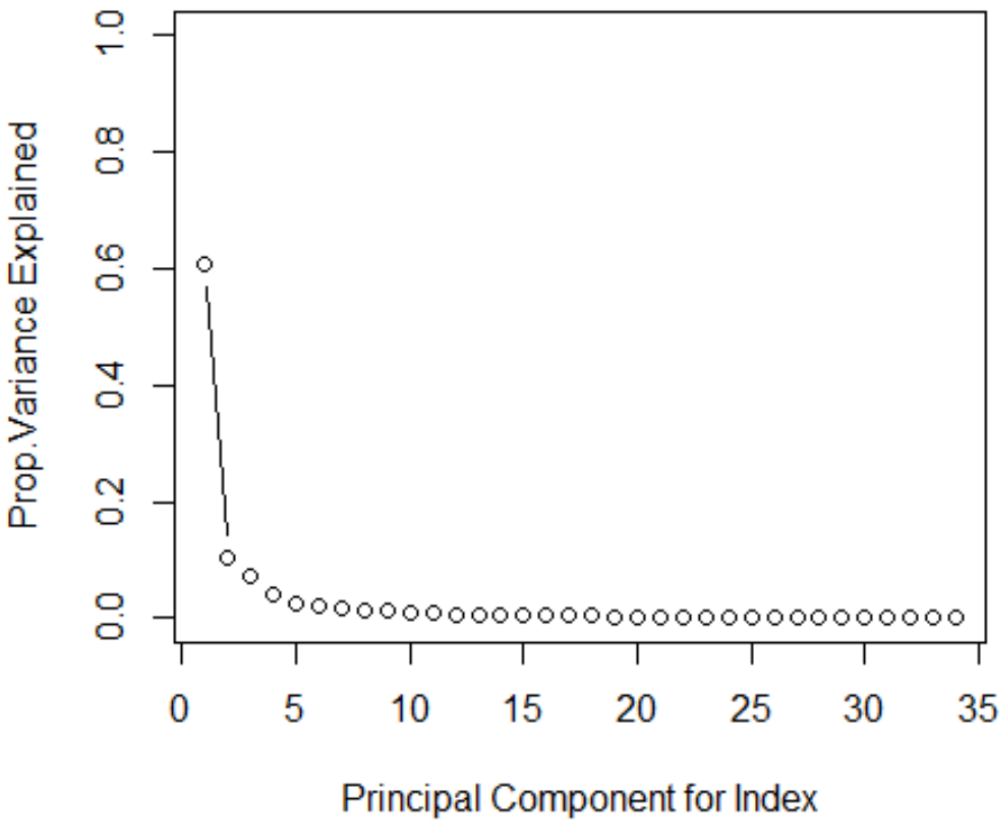
share (EPS), revenue, net income and other related variables as control variables. We extract factors from these control variables representing the company's achievements and operations by using the same method depicted earlier in the study. Our data set contains weekly data starting from June 2013. We also obtain factors related to Tencent by employing correlation coefficients and PCA. To improve the prediction, we adopt two correlation methods to calculate correlation coefficients, namely, the Pearson correlation and the lead-lag correlation methods. After computing the correlation, we apply PCA to extract the factors. The critical value of correlation is 0.4. Table 17 shows the results of lead-lag correlation methods; 43 keywords are observed to be strongly leading-correlated with 0700.HK. PCA assesses the explanatory performance of factors, with results presented in Figures 6. The prediction results of different methods are displayed in Figures 7 and 8. We compare the performance of prediction from these two methods by using the mean absolute error (MAE) and mean absolute percentage error (MAPE). The results are presented in Table 18. Lead-lag correlation methods may perform slightly better than the Pearson correlation method, given that both error statistics of the former are smaller than the corresponding values of the Pearson correlation method.

	Keyword		
1	腾讯股价(Tencent stock price)	27	pubg mobile
2	腾讯市值(Tencent market value)	28	amazon stock price
3	腾讯股票(Tencent stock)	29	snap stock
4	腾讯财报(Financial report of Tencent)	30	tcehy stock price
5	马化腾(Pony)	31	10 cent
6	0700.HK	32	who owns epic games
7	股王(King of stock)	33	arena of valor
8	腾讯控股(Tencent Holdings)	34	hang seng index
9	微信(WeChat)	35	who owns fortnite
10	QQ	36	tencent ir
11	wechat	37	djia today
12	吃鸡(PlayerUnknown's Battlegrounds)	38	tencent epic games
13	王者荣耀(King of Glory)	39	tencent share price
14	微信支付(WeChat pay)	40	tencent us stock
15	腾讯视频(QQ video)	41	jd stock
16	腾讯手游(Tencent mobile phone games)	42	阿里云(Alibaba Cloud)
17	腾讯(Tencent)	43	腾讯手游(Tencent mobile phone games)
18	tencent	44	腾讯（繁体）(Tencent(in traditional Chinese))
19	tencent pubg	45	腾讯股价（繁体）(Tencent stock price(in traditional Chinese))
20	pubg	46	腾讯创造101（繁体）(Tencent Produce 101 (in traditional Chinese))
21	fortnite	47	腾讯股价走势（繁体）(Tencent stock price trend (in traditional Chinese)
22	epic games stock	48	腾讯诛仙（繁体）(Tencent Jade Dynasty (in traditional Chinese))
23	tencent stock price today	49	魔力宝贝腾讯（繁体）(Cross Gate Tencent (in traditional Chinese))
24	fb stock	50	恒生指数（繁体）(Hang Seng Index(in traditional Chinese))
25	baba stock price	51	dow jones index
26	tencent games pubg	52	700

Table 16: Keywords selected for Tencent

	$\rho > 0.4, lag >= 0$	$\rho > 0.5, lag >= 0$	$\rho > 0.6, lag >= 0.6$
	腾讯股价(Tencent stock price)	腾讯股价(Tencent stock price)	腾讯股价(Tencent stock price)
	腾讯股票(Tencent stock)	0700.HK	0700.HK
	0700.HK	QQ	吃鸡(PlayerUnknown's Battlegrounds)
	QQ	wechat	微信支付(WeChat pay)
	wechat	吃鸡(PlayerUnknown's Battlegrounds)	腾讯手游（Tencent phone game)
	吃鸡(PlayerUnknown's Battlegrounds)	微信支付(WeChat Pay)	tencent
	腾讯手游(Tencent mobile phone games)	tencent	fortnite
	tencent	pubg	epic games stock
	tencent pubg	fortnite	tencent stock price today
	pubg	epic games stock	baba stock price
	fortnite	tencent stock price today	amazon stock price
	epic games stock	fb stock	tcehy stock price
	tencent stock price today	baba stock price	arena of valor
	fb stock	pubg mobile	hang seng index
	baba stock price	amazon stock price	who owns fortnite
	tencent games pubg	tcehy stock price	djia today
	pubg mobile	who owns epic games	tencent epic games
	amazon stock price	arena of valor	tencent share price
	tcehy stock price	hang seng index	tencent us stock
	who owns epic games	who owns fortnite	腾讯手游(Tencent phone game)
	arena of valor	djia today	腾讯（繁体）(Tencent(in Traditional Chinese)
	hang seng index	tencent epic games	腾讯股价（繁体）(Tencent stock price(in Traditional Chinese)
	who owns fortnite	tencent share price	恒生指数（繁体）(Hang Seng Index(in Traditional Chinese))
	djia today	tencent us stock	dow jones index
	.....	.....	
Total	43	31	25

Table 17: Strongly correlated and leading keywords obtained by lead-lag correlation methods



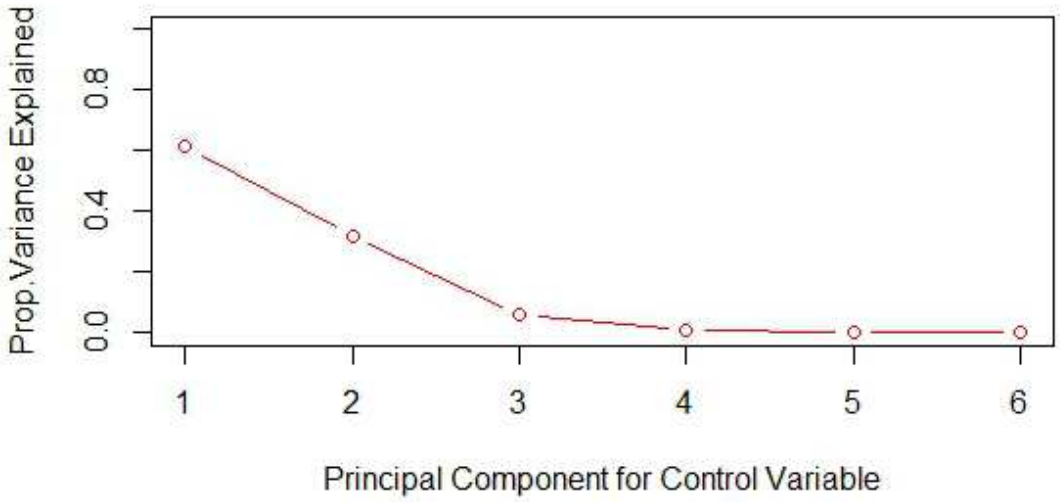


Figure 6: Explained variance of PCA using lead-lag correlation methods

	MAE1	MAE2	MAPE1	MAPE2
1	1.6199	1.5635	0.0038	0.0037

Table 18: Prediction performance of two correlation methods

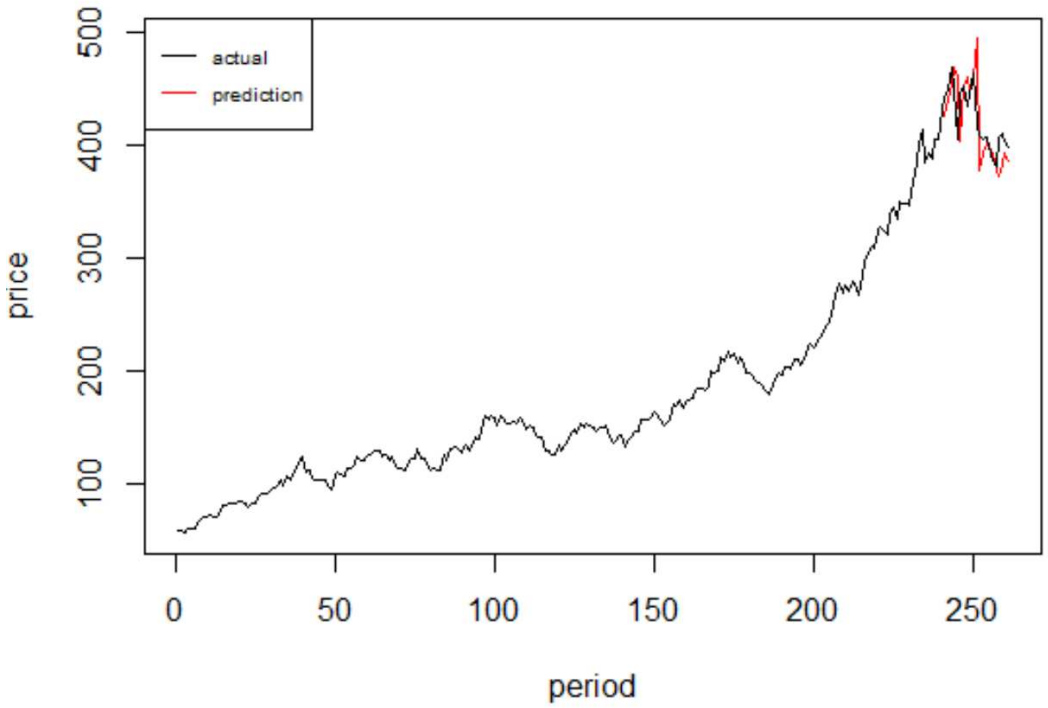


Figure 7: Stock price forecast using the lead-lag correlation method

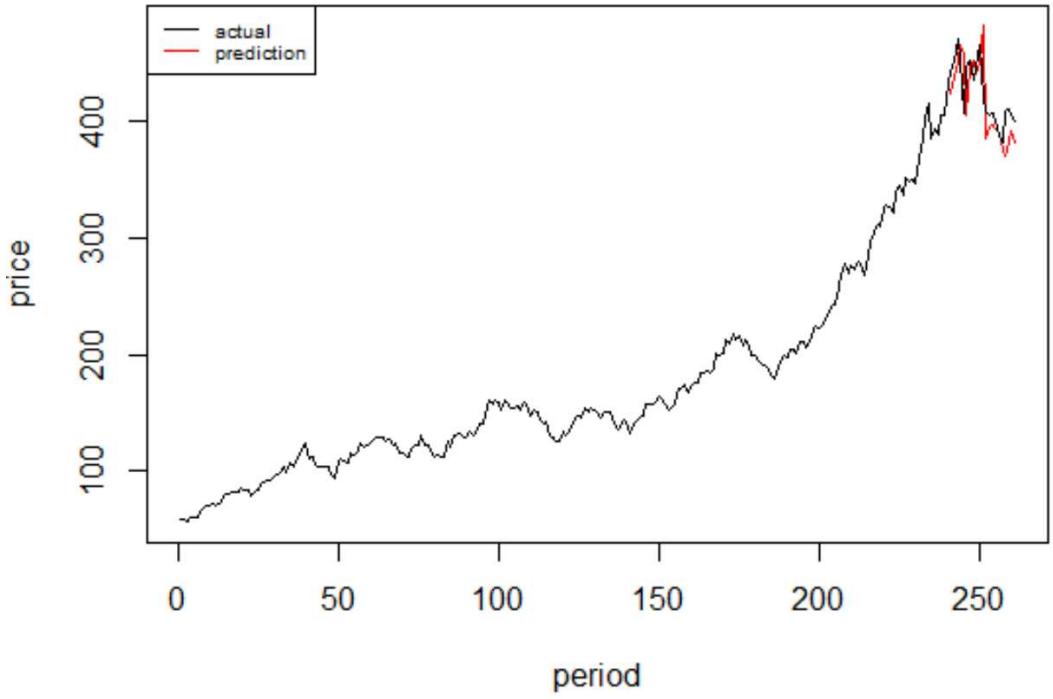


Figure 8: Stock price forecast using the Pearson correlation method

7 Conclusion

This paper analyzes the effects of search indices on the Chinese stock market. We initially utilize the Pearson correlation coefficient to find the correlation between search indices and the A-share market. The results reveal a strong correlation between the Shanghai Composite Index and the keyword series. We classify the keywords into four types where each type represents a different aspect of investors' sentiment. Different factors are extracted from these keyword series and represent different types of investors' sentiment. The regression results of the FAVAR model indicate that investors' sentiment and the Shanghai Composite Index have significant effects on each other. Changes in investors' sentiment cause fluctuations of stock price; a shock to composite stock returns as well causes a notable fluctuation in investors' sentiment. The result that investors' sentiment in the trading stage has significant effects on the return of stock price suggests that the information investors obtain from the Internet may influence their trading strategies. In addition, the economy and government policy are the general factors that grasp investors' attention and influence their investment decisions. We can observe from the results that investors' attention to the economy and policy significantly affects the return of the Shanghai Composite Index. Changes in stock returns also have significant influence on the four aforementioned types of investors' sentiment.

The impulse responses demonstrate that stock returns can exhibit fluctuations as a result of shocks to factors, and vice versa. Generally, such

fluctuations continue for a few days and then cease. As we do not categorize the keywords as positive or negative ones, we cannot simply interpret that a positive shock to investors' sentiment can cause a positive or negative change in stock return. Both negative news and economic improvements can increase search frequency but have different impacts on stock returns. The effects of variables on stock price returns appear to be less prominent than those signified by other regressions. One explanation is that, with daily data, a shock to investors' search frequency in one day does not play a material role in affecting the composite stock price of the financial market because of the large size of the market.

In the second part, the keywords which have a strong relationship with the Hang Seng Index within the two distinct periods can be dissimilar. The main difference is evidently due to the Shanghai-Hong Kong Stock Connect. Tables 13–15 present the Pearson correlation coefficients for the relationship between the Hang Seng Index and the keywords in different periods, from which we obtain three findings. First, the number of keywords that are strongly correlated with the Hang Seng Index is smaller than that with the Shanghai Composite Index because our keyword series are chosen manually, thereby introducing subjectivity into data selection. Most keywords are recommended by the Baidu Index, which means that they are selected on the basis of their significance to Chinese investors. While many Chinese companies participate in the Hong Kong financial market, certain foreign

companies' stocks also appear, but these may not be closely related to our selected keywords. Second, we observe different keyword types influencing the Hang Seng Index during different periods, as shown in Tables 13–15. Investors' sentiment on trading and the economy are the general influential factors affecting the Hang Seng Index. Tables 13–15 confirm that keywords relevant to the Shanghai-Hong Kong Stock Connect are obviously correlated with the Hang Seng Index in the second period.

We can thus conclude that investors pay much attention to the impact of the Shanghai-Hong Kong Stock Connect on the Hong Kong financial market. Third, if we compare the coefficients of related keywords for the Shanghai Composite Index and the Hang Seng Index, we can infer that investors' in the A-share market focus on the action concerning the Southbound trading, whereas those in the Hong Kong financial market pay attention to that of the Northbound trading. This interesting phenomenon can be reserved for future discussion.

To illustrate the predictive power of investors' sentiment, this paper constructs a FAVAR model which incorporates investors' sentiment factors to predict the stock prices of Tencent. For comparison, we employ lead-lag and Pearson correlation coefficients to construct the model. The results of the two approaches, as displayed in Figures 7 and 8, are not significantly different.

References

[1]. Bernanke B S, Boivin J, Eliasz P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. The Quarterly Journal of Economics, 120(1): 387-422.

[2]. Brynjolfsson E, Hu Y J, Rahman M S. (2013). Competing in the age of omnichannel retailing. MIT Sloan Management Review, 54(4): 23-29.

[3]. Challet D, Ayed A B H. (2014). Do Google Trend data contain more predictability than price returns?. arXiv preprint arXiv:1403.1715.

[4]. Choi H, Varian H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1): 2-9.

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

[6]. Fernald J G, Spiegel M M, Swanson E T. (2014). Monetary policy effectiveness in China: Evidence from a FAVAR model. Journal of International Money and Finance, 49: 83-103.

[7]. He Q, Leung P H, Chong T T L. (2013). Factor-augmented VAR analysis

of the monetary policy in China. China Economic Review, 25: 88-104.

[8]. Li H, Yang S. (2013). Application of Linear Regression Analysis Model in Stock Investment. Mathematical Computation, 2(2): 36-39.

[9]. Liu L X, Shu H, Wei K C J. (2017). The impacts of political uncertainty on asset prices: Evidence from the Bo scandal in China. Journal of Financial Economics, 125(2): 286-310.

[10]. Luong T A, Shi C M, Wang Z. (2019). The impact of media on trade: Evidence from the 2008 China milk contamination scandal. Available at SSRN 3164244.

[11]. Preis T, Moat H S, Stanley H E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific Reports, 3(1684): 1-6.

[12]. Vosen S, Schmidt T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. Journal of Forecasting, 30(6): 565-578.

[13]. Wu L, Brynjolfsson E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales//Economic analysis of the digital economy. University of Chicago Press, 89-118.

[14]. Xu S Y. (2014). Stock price forecasting using information from Yahoo Finance and Google trend. UC Berkeley.

[15]. Liu Y, Lv B, Peng G (2011). Predictive power of Internet search data for stock market: A theoretical analysis and empirical test. Economic Management, 33(1):172-180.