THE IMPACT OF SOCIAL MEDIA SENTIMENT ON CHINESE STOCK MARKET

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Applied Economics and Management

by

Ge Yu

May 2021

ABSTRACT

Using the newest data from a leading Chinese social media investing platform, I test the impact of the individual and institutional investor sentiment on stock market using VAR models. Although my results show that the investor sentiment can influence stock market, their impact is different: individual investor's sentiment is helpful in predicting the stock return while institutional investor's sentiment is helpful in predicting the trading volume. In addition, I find that the firm size can influence the impact of investor sentiment: for larger firms, the impact of sentiment is more obvious. My study suggests that it is meaningful to split the individual and institutional investor sentiment, and it is necessary to choose a balanced stock portfolio while studying the impact of investor sentiment.

BIOGRAPHICAL SKETCH

Ge Yu is a Master student in Charles H. Dyson School of Applied Economics and Management at Cornell University. He got his Bachelor of Economics from Central University of Finance and Economics in Beijing. His research interests are behavior finance, asset pricing and experimental economics. His email address is gy88@cornell.edu.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Professor Jawad M. Addoum for his patient guidance and generous help in my graduate study, research and my Ph.D. application. His support helps me to choose an interesting topic to start my master thesis. His experience in research guides me to successfully move on in each step. I would also like to thank Professor Calum Turvey. As a senior faculty in agriculture finance, his suggestions are essential for my study and my thesis.

And my parents, I want to thank them for their supports especially during the pandemic. Their encouragement is always the source of my valour. Also my classmates and friends, many thanks for their help and company in this two year journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In history, the prediction of stock market is usually based on historical data and financial data of the stock itself. This kind of prediction model clearly neglects the investors' behavior in stock markets. In 1990, the DSSW model introduces the concept of noise trader to asset pricing theories (De Long et al., 1990). After that, investors sentiment become a hot topic in both traditional and empirical asset pricing studies in stock market. In these studies, investors sentiment is usually defined as a false belief or an irrational expectation on stock returns (Baker and Wurgler, 2006; Barberis et al., 1998; Shefrin and Belotti, 2008). The tradition ways to evaluate investor sentiment in empirical studies are usually by surveys from investors or consumers and compositing index based on frequently-used variation in various principal proxies for sentiment (Baker and Wurgler, 2006, 2007; Baker et al., 2012). With the development and ubiquity of Internet, studies start to use online information to build proxies for investor sentiment. These proxies are commonly combined with natural language processing and sentiment analysis techniques. However, in previous literature, former researchers are often too focused on institutional investors' opinions such as news and analysis reports. As a consequence, they usually neglect individual investors' views on stock market.

Social media provides a marvelous opportunity to study both institutional investors and individual investors' sentiment. As a platform with both read and write attributes, investors can exchange their views on the stock market and read updated news (Bukovina, 2016), such as the world leading trading platform StockTwits. But will investors really trade based on Internet post? A

recent experimental study says yes. The experiment shows that there indeed exists investors who will use the online forums as references and make trading decision (Ammann and Schaub, 2020). This study further validates the impact of social media sentiment. Researchers also notice the power of social media and study its impact on stock market (Kumar and Lee, 2006; Das and Chen, 2007; Joseph et al., 2011; Shi et al., 2017). Nevertheless, many of these studies such as (Baker et al., 2012; Shi et al., 2017) only consider the stock return or stock price, but neglect other variables of stock market, like trading volume or volatility. Also, their stock portfolios are typically biased towards large companies (Baker et al., 2012; Shi et al., 2017). Therefore, their conclusions are not representative for all kinds of stocks.

In this thesis, I choose a leading Chinese investing forum as the targeted social media platform. After choosing a more balanced stock portfolio, I use an advance pre-trained sentiment analysis model to transform the posts to investor sentiment index. By splitting the sentiment index into individual and institutional investors' sentiment, I study both the "read" and "write" attribute impact on stock market, including the stock price, stock return, trading volume and volatility. Using the sentiment index, I find evidence that investor sentiment can indeed influence stock market. For individual investors' sentiment, it can help predicting the stock return. For institutional investors' sentiment, it can help predicting the trading volume. After that, I also study sentiment with the heterogeneous beliefs and firm sizes, which have not been discussed in previous investor sentiment research. I find no evidence that heterogeneous beliefs can influence stock market. But I can observe significant difference in the investor sentiment impact among large size firms and small size firms.

My thesis contributes to the literature in the several following ways. First, my findings provide clear evidence that investor sentiment can impact the stock market. Second, I find evidence that the firm size can influence the impact of investor sentiment. Third, my data set provides opportunity to relate the financial regulation to investor sentiment. Further research can be focused on the impact of regulation such as limit to arbitrage or the price limit.

The following parts of this thesis are organized as follows. The second chapter is the literature review, where I review theoretical and empirical studies of investors sentiment and stock market. I also review technical progress of sentiment analysis in this part. The third chapter is the data part. In this part, I explain in detail how I obtain the data, how I process the data and the descriptive statistics. I also brief introduce the working principle of the pre-trained sentiment analysis model which using here. The fourth chapter is the methods part, where I introduce the theoretical basis and the econometrics setting. The fifth chapter is the cardinal part of this thesis. In this chapter, I show the results of our empirical econometrics models, and interpret these results. I also expand my topics to heterogeneous beliefs and firm size later in the discussions part of this chapter. In the last chapter, I summarize all my findings, and provide suggestions for possible future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Investor Sentiment and Stock Market

According to the definition of Shiller, behavior finance is a research area where wilder social science perspectives, such as psychology and sociology, are introduced into financial models to handle the market anomalies problem (Shiller, 2003). Among the classical behavior finance theories, the majorities are overconfidence theory, prospect theory and loss aversion. Based on these classical theories, behavior finance has drawn many researcher's focus and commitments.

In the concepts of studies in behavior finance, investor sentiment and its relationship with stock market acquires much attention (Barberis et al., 1998; López-Cabarcos et al., 2020). According to the analysis of López-Cabarcos et al., the reason of increase in studies in investor sentiment might be the emergence of a new measure of sentiment encouraged by social media and the popularity of natural language processing (NLP) models (López-Cabarcos et al., 2020). Previous research papers have proposed several different kinds of definitions of investor sentiment. For instance, a classical paper by Barberis et al. believed that investor sentiment is the way "how investors form beliefs" (Barberis et al., 1998). In 2006, Baker and Wurgler defined investor sentiment as "the propensity to speculate", indicating that the investor sentiment is investor's optimism or pessimism about the investments in the future (Baker and Wurgler, 2006). From these definitions, several divergent methods are proposed to measure the investor sentiment. Basically, these methods can be divided into three kinds: surveys from investors or consumers, composite index based on frequently-used

variation in various principal proxies for sentiment such as Baker and Wurgler's research (Baker and Wurgler, 2006, 2007) and the further research by Huang et al. (Huang et al., 2015), and the presently popular textual analysis from traditional and modern social media (McGurk et al., 2020). These methods provide a solid empirical foundation to construct a new measurement to evaluate the investor sentiment.

Based on the above methods, researchers tried to find the possible correlation between investor sentiment and the stock market: especially the cross section and future returns. In 2000, Fisher and Statman categorized the investor sentiment into Wall Street strategists, individual investors, and newsletter writers three groups. They showed that there was a negative relationship between the sentiment of the three groups and future stock returns (Fisher and Statman, 2000). On the contrary, in 2004, Brown et al. found that sentiment had little predictive power for near-term future stock returns. They also cannot find evidence that sentiment primarily affects individual investors and small stocks (Brown and Cliff, 2004). After that, they used sentiment indexes from surveys (so-called accurate measure of investor sentiment) and concluded that irrational sentiments did affect asset prices and asset pricing models should take sentiments into account (Brown and Cliff, 2005). In 2006, Kumar and Lee used a large database of retail investor transactions over 1991-1996 and they found evidence supporting the role for investor sentiment in financial markets (Kumar and Lee, 2006). In line with the above research for the US stock market, Schmeling wielded consumer confidence as a proxy for investor sentiment from 18 industrialized countries and found similar evidence that sentiment negatively impacted aggregate stock returns across countries on average. A cross-section regression also found that this impact is stronger in countries with less market

integrity and likely to herd behavior in culture (Schmeling, 2009). Similar to this research, a supplementary research built six sentiment indices from six developed countries and decomposed these indices into "global" and "local" sentiment. The following investigation found homogeneous evidence that sentiment had negative influence on small, high return volatility, growth, and distressed stocks for local sentiment. They also found that total sentiment was an opposing predictor for market returns in country-level (Baker et al., 2012).

Present researchers tend to measure investor sentiment from traditional and modern social media due to the development of Internet, textual analysis techniques and sentiment analysis models. Using the textual analysis techniques: in 2007, Das and Chen constructed the sentiment index from stock message boards in Yahoo! and observed proofs for the relationship between index levels, volumes, volatility and postings in value-tech-sector (Das and Chen, 2007). In 2011, Joseph et al. used online ticker searches and found that when an increase in search density occurs, it usually forecasts abnormal stock returns and extravagant trading volumes (Joseph et al., 2011). Comparably, in 2015, Da et al. also utilized household daily Internet search volumes and composed a new measure of investor sentiment. They gave proofs that this index can help predict return reversals in short-term, help predict excessive trading volatility, and help predict the activity of mutual fund: flowing out of equity and into bonds (Da et al., 2015). One recent research also utilized a large dataset on StockTwits and tested different machine learning algorithms. Their investigation showed that although there is correlation in sentiment and stock returns, the sentiment constructed from the messages from social media cannot help predict large capitalization stock returns at a daily frequency (Renault, 2020).

In Chinese literature, investor sentiment and stock markets are also a hot topic recently. Generally speaking, most of the research used equivalent methods and theories. In 2007, the regression model of Q. Zhang et al. showed that the institutional investors's sentiment has influence on the stock price, but no evidence showed that the individual investor's sentiment has a relationship with the market (Zhang et al., 2007). Their further research was based on the noise trading theory (De Long et al., 1990). Their results showed that stock prices vary with the volatility of investor sentiment and more importantly, the influence of positive change is stronger than that of passive change (Zhang and Yang, 2009). Similar to Baker and Wurgler's research (Baker and Wurgler, 2006, 2007), Jiang and Wang used principal components analysis to constructed their own sentiment proxies. Using their proxies, the model suggested that investor sentiment has significant aggregate effects and cross-section effects on stock returns (Jiang and Wang, 2010). This conclusion is nearly the same as other research in US. Just like other non-Chinese literature, Chinese researchers were also interested in Internet tools and machine learning tools. For instance, in 2014, Y. Zhang et al. adopted the internet search data and validated that the intensity can help predict the stock's short-term return, trading volume and cumulative return. They also discovered that the search intensity can work as a better proxy for traditional investor sentiment index (Zhang et al., 2014). Chinese researchers also noticed that the popularity of social media and stock trading platforms. In 2015, Yi et al. conducted a word frequency statistic to the posting of a trading BBS and constructed a VAR model. They discover that bearish sentiment can cause higher trading volumes but bullish sentiment can have a significantly higher impact on stock prices (Yi et al., 2015). In 2017, Shi et al. utilized the sentiment index and attention index provided by Xueqiu and Uqer

and provided evidence that individual investors' attention index has great influence on Chinese stock market. They cannot find proofs that news sentiment index has obvious relationship with stock market (Shi et al., 2017).

Overall, contemporary researchers tend to combine latest techniques with behavior finance theories, to construct an accurate, reliable, and interpretable proxy for investor sentiment. Nevertheless, although there have been many papers focusing on the social media, most researchers were keener on the traditional way to construct the sentiment proxy. After the construction of proxies, researchers try to answer the questions such as whether investor sentiment will impact stock market, how investor sentiment impacts stock market, what parts and what variables in stock markets might be influenced by investor sentiment, sentiment of which kind of investors is more prone to influence the stock market, etc. From English and Chinese literature, although there exist different voices, popular opinion believes that sentiment plays a significant role in stock markets. Despite the argues among some papers, these research projects provide opportunities to compare, test and verify the results of this thesis.

2.2 Sentiment Analysis Development

The origin of sentiment analysis can be found in a study in methods to evaluate public opinion in the early 20th century (Mantyla et al., 2018). But due to the technical limits, most research on the sentiment analysis based on Internet texts are published after 2004 (Mantyla et al., 2018). Among the sentiment analysis techniques, the most frequently used are natural language processing (NLP) and machine learning algorithms. In 2002, Bo Pang, Lillian Lee and Shivakumar

Vaithyanathan employed three supervised machine learning algorithms (Naive Bayes, Maximum Entropy Classification, and Support Vector Machines) to perform sentiment classification tasks based on movie reviews data (Pang et al., 2002). Similarly, Turney (Turney, 2002) presented an unsupervised sentiment classification model based on travel reviews. These papers are usually considered as the beginning of modern NLP and sentiment analysis models. With the development of NLP programming, there also exists many sentiment analysis model using lexicon-based methods, which are constructed by utilizing a list or dictionary of opinion words (Ding et al., 2008). For instance, Taboada et al. (Taboada et al., 2011) built the Semantic Orientation CALculator (SO-CAL), which was based on dictionaries of words and phrases annotated with their polarity and strength.

With the popularity of Internet and deep learning algorithms, papers begun to present more advanced models and use bigger datasets to improve the performance of sentiment analysis. In 2013, the Sentiment Treebank presented by Socher et al. demonstrated that the Recursive Neural Tensor Network and deep learning methods can help improve the accuracy of sentiment analysis (Socher et al., 2013). Based on that, researchers started to implement similar methods to enhance the performance of sentiment analysis models. In 2014, Kalchbrenner et al. introduced a Dynamic Convolutional Neural Network (DCNN) model. Their later tests proved that the model achieved outstanding performance in most of the sentiment analysis situations (Kalchbrenner et al., 2014).

Meanwhile, because of the wilder application of social media in commercial and private sectors, these platforms such as Twitter have attracted much attention. In 2016, Saif et al. proposed an approach allowing for detecting sen-

timents from both entity-level and tweet-level (Saif et al., 2016). Their approach outperformed the baselines in most of the datasets from Twitter. In fact, in 2020, a survey (Adwan et al., 2020) showed that recently there were over 40 different approaches to Twitter sentiment analysis. These approaches can be roughly identified into four categories: machine learning, lexicon-based, hybrid-based, and graph-based.

The above research illustrated a fast evolution in English sentiment analysis approaches. Some of these approaches produced powerful open source packages in analyzing English texts, such as NLTK in Python (Bird et al., 2009) or TM module in R (Feinerer et al., 2008). Nevertheless, the evolution of research in Chinese sentiment analysis remarkably fell behind the research in sentiment analysis in English textual. Unlike English, or other similar European languages, Chinese texts need a specific segmentation step to transform the raw data (Peng et al., 2017). Currently there are three major segmentors: ICTCLAS invented by Dr. Zhang Huaping (Zhang et al., 2003), THULAC developed by Tsinghua University (Li and Sun, 2009) and an open source package Jieba on GitHub. After segmenting the characters, researchers usually applied three kinds of approaches similar to English texts: machine learning based, knowledge based and mix models (Peng et al., 2017). These models are known as Monolingual Approaches. Symmetrically, Multilingual Approaches are also frequently used in Chinese sentiment analysis. The key idea of these approaches is to take advantage of the existing resources in English sentiment analysis by translating the Chinese texts into English or translating the English resources into Chinese. For instance, in 2008, Wan prosposed an unsupervised model by fully using bilingual knowledge (Wan, 2008). He found that the first method is better than the second one, and their combination can further improve the per-

formance of the model. In 2015, to reduce the negative influence of the translation between Chinese and English, Chen et al. proposed a novel boosting knowledge validation model (Chen et al., 2015). Their experiments proved that the model is efficient and more importantly, necessary.

To conclude, sentiment analysis is always a hot topic in both English and Chinese research. Especially in English, the sentiment analysis has over 70 years of history (Cambria and White, 2014). Although the development of Chinese sentiment analysis fell behind at the beginning, more and more researchers are concentrating on this field. The devotion of researchers in this area has guaranteed that the sentiment analysis techniques are reliable and valid for further financial research.

CHAPTER 3

DATA

3.1 Textual Data Collection

Technically, social media are often defined as high interactive platforms which are based on web or mobile technologies (Kietzmann et al., 2011). One major product provided by social media platforms is the so-called "big data". Because of the high-volume, high-velocity and high variety information, we cannot process the big data via traditional data management and analyzing tools. And it is not unusual for finance research to apply social media big data in models. Focusing on the impact of social media sentiment on Chinese stock market and investor behavior, here, I thrive to acquire data from social media and stock market. This part briefly introduces how the textual data from social media platforms is collected and how the textual data is categorized. According to the review of Bukovina, social media platforms usually have two attributes (Bukovina, 2016):

(1) "Read". A typical social media platform can provide users useful information which can be utilized by users to make investing decisions. This information can be created by other media, the platform itself, and most importantly, other users. This means, the information on the social media platform can be divided as two kinds: "news", which is created by professional institutions or analysts; and "comments", which is the posting written by other users.

(2) "Write". Social media usually offers users a platform to communicate with other users. A traditionally Chinese financial social media platform is usu-

ally in a type of a stock investing forum. In such a stock investing forum, there are many subforums focusing on different stocks. In these subforums, users can post their comments about the stock. These comments generally contain their predictions for the price change, the turnover rate, and the trading volume. Sometimes, these comments also contain their predictions for other stocks, the whole stock market, or the comments of hotspot topics. These subforums also have news and professional stock market analysts' comments from traditional media.

Initially, to analysis the investor sentiment of social media, I need to choose a social media platform to acquire the postings of users and the financial news from professional institutions. Currently, there are many popular stock investing forums in China run by either web technology company or professional investment consulting companies, such as Sina Finance, Guba, Tencent Finance, etc. This research chooses Guba (literally meaning "stock forum") of Eastmoney as the target financial social media platform. The reasons are:

(1) It is one of the leading stock investing forums in China. The large number of registered users and their frequent activities can provide us enough textual data of comments and news. According to the data from Dongguan Securities Co. Ltd. (2020), Eastmoney serves over 60 million users monthly. As an important product of Eastmoney since June 2006, Guba is believed to have strong user stickiness because of its various contents and unrestrained communication between individual investors.

(2) Previous Chinese research also chose this investing platform such as the work of Shi et al. (Shi et al., 2017). The previous research projects have discovered evidence of the possible correlation between investor sentiment and stock

market. Therefore, it might help to better compare the results of this thesis to other previous findings. Meanwhile, these research projects also demonstrate the rationality of choosing Guba as the target financial social media platform.

(3) It is relatively easier to write a web crawler regarding this website. The html structure of this website is relatively simple and can be readily decoded. Technically, it might help to improve the response rate of the web crawler program and largely cut the working time. This is meaningful to build a higher quality dataset more efficiently.

After choosing the target social media platform, the next step is to analyze the structure of this social media platform Guba. Like other popular investing forums, Guba is consisted of different kinds of subforums. These subforums can be roughly categorized as the subforum for specific individual stock, and the subforum for specific topic. The subforum for specific individual stock is designated for discussion on one specific stock, usually named by the name of the company and the stock code. In this kind of subforum, the posts are typically from individual users and institutional users. Individual users normally post their own comments and prediction for this specific stock, while institutional users generally write newsletters, professional analysis for the stock or announcement from the company or authorities. A convenient way to separate the posts from individual users and institutional users is to utilize the "read time" statistic. Here, I randomly choose a small part of the data set, and manually label the posts as individual users' post and institutional users' post. I find that typically, the read times of posts from individual users is below 3000 and the read times of posts from institutional users is regularly over 3000 (and higher at most times). Therefore, by establishing this threshold on read time, we

can accurately separate the posts from individual and institutional users. The subforum for specific topic is designated for discussion on certain topic, usually named by hot topics or certain commodities. In this kind of subforum, the posts are typically focused on the specific topic. For example, for the subforum named "financial comments", the posts are basically financial news, financial analysis or investing suggestions written by professional analysts. For the subforum name "gold forum", the posts are all about the trading of gold from both individual and institutional users. Therefore, as we can see, it is quite hard to separate posts from individual and institutional users because we cannot make sure whether a subforum for specific topic is designed for institutional posts. Hence, this thesis only focuses on the first kind of subforum – for a specific stock. After choosing the stocks to study, I can confirm the subforums to collect textual data. This will be discussed in the next part.

Given the characteristic of big data, collecting textual data from the target subforums is not an easy job. In fact, for a frequently discussed stock, it will need to browse over 500 pages of posts, which is over 50,000 posts in total, to collect textual data for half a year. This determines that I should build an automatic data collecting program. This kind of program is known as a web crawler, or sometimes called web scraper, spider or spiderbot. One important task in building a web crawler is to parse HTML and XML documents scraped from the web crawler. The HTML parser this thesis uses here is Beautiful Soup in Python. The output of the web crawler is a long table with four columns: respectively read times, comment times, posting contents and time.

After getting the final results table, one final step is to separate the posts into individual users' posts and institutional users' posts. A simple approach

is, as stated before, to adapt the threshold 3000 read times. Any post whose read times is below 3000 is categorized as individual users' post; and any post whose read times is above or equal 3000 is categorized as institutional users' post. According to my test, this threshold can efficiently and accurately separate the data.

However, this method only considers institutional users' posts posted on the particular subforum. In fact, institutional users' posts such as newsletter and analysis reports can influence the stock's industry even though they are not directly related. This means that this method will neglect the impact from external factors. There is another problem found when I segment the comments: not every day each subforum has institutional users' posts. This causes missing data in various days. Hence, this thesis decides to bypass this problem by using the news sentiment data provided by Uqer, a famous quant trading platform in China. Previous researchers have demonstrated the reliability of data on Uqer, such as Shi et al. (Shi et al., 2017). The news sentiment data provided by Uqer is composed by the stock code, the published date of news and report, the score of related and the sentiment score of the news/report. Then we can calculate the sentiment score of each news/report weighted by the score of relation.

3.2 Stocks Selection

The previous part decided to choose Guba's subforum for specific stock as the target social media platform. To further continue the data collecting work, we need to determine the stocks to investigate. This thesis uses the term "stock portfolio" to represent the collection of different kinds of stocks to gather posts

and stock market related data.

Previous research work provided some clues of how to construct a "proper" stock portfolio. The paper of Shi et al. used an attention and sentiment index composed by all stocks in CSI 300 Index, which is capitalization-weighted and designed to mimic the performance of the leading 300 stocks on Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE) (Shi et al., 2017). Therefore, their stock portfolio should be all these stocks. However, their research did not compose their own indices. Instead, the data they used here is provided by Uqer, a leading trade consulting platform in China. Hence, we do not know how the attention and sentiment index are exactly compounded. The second problem in this stock portfolio is that it is overwhelming biased towards leading companies with large capitalization. According to the data provided by Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE), the average circulation market value per stock of SSE is 23.48 billion RMB and of SZSE is 10.94 billion RMB. Nevertheless, the average circulation market value per stock of stocks in CSI 300 is over 134.57 billion RMB, which is nearly 6 times to SSE and 12 times to SZSE[1]. Another problem of this stock portfolio might be the old approval-based initial public offering (IPO) system, through which all the 300 stocks went public. Under this IPO system, companies are strictly screened by the authority and the pace of issuance of stocks is rigorously controlled. Companies demonstrate their ability for a sustainable profitability to qualify for a formal approval from the China Securities Regulatory Commission (CSRC). Furthermore, this system also puts regulation on the IPO pricing. Hence, it is possible that some investors will tend to be more optimistic about the stocks since they all went through rigor financial censor and auditing. Mean-

[1]Data on February 26th, 2021.

while, because of the opaque approving process, some investors might tend to be more pessimistic about the stocks. In fact, there has been evidence that Chinese approval-based IPO system might impact both individual and institutional investors' sentiment (Li, 2018).

Chinese policymakers and CSRC also notice the deficits of this old approval-based IPO system, but it is a long way to completely reform. Although a new registration-based IPO system was promoted back in 2013, it was not launched in relatively small scale until 2020. Under this new, international-standard registration-based IPO system, enterprises no longer need to wait for approval. Instead, like US or other developing countries, they now only need to disclose certain information to investors to publicly list their shares. Even though this reform is limited in STAR Market of SSE (Official name: Shanghai Stock Exchange Science and Technology Innovation Board) and ChiNext of SZSE, both of which are platforms designed for "implementing the national strategy of independent innovation", this reform still provided us an opportunity to construct a reasonable stock portfolio.

On August 24th, 2020, the first 18 companies successfully went public on ChiNext through this pilot registration-based IPO system. This pilot system cancelled the regulation on daily volatility for the first five days after IPO. It also eased the original 10% restriction on volatility to 20%. This thesis chooses these 18 companies to construct the stock portfolio because of the registration-based IPO system. The second important reason to choose the 18 companies is that they are from various different industries, and the value of their stock have tremendous difference. So, this stock portfolio can represent numerous types of enterprises. One of the disadvantages of the stock portfolio is that individual

Table 3.1: Categorizing Table of the Stocks

Scale	Industries	Stock Codes
Large	Media	300860
	Material	300861,300875
	Electronic	300866
	Environment	300867
	Medical	300869
	Internet	300872
Small	Material	300865,300868,300876,300877
	Electronic	300862,300870
	Environment	300864
	Medical	300871,300878
	Vehicle	300863
	Transport	300873

This table presents the stock categorizing result. In this table, I define a stock/company is large if its total market value at the end of IPO day is equal to or over 10 billion RMB.

investors should qualify to invest in ChiNext. Some of the requirements are over 24 months stock investing experience and over 100,000 RMB total assets. This might filter some individual investors and noise trader (De Long et al., 1990). Another disadvantage is that the length of the trading data is limited, due to the short time.

Basically, the stocks in the stock portfolio can be categorized as in the Table 3.1.

Now I can successfully gather the posts textual data. By using Jieba aforementioned, I can split the posts into single Chinese words. Here, I summarize the top 20 frequently used words/phrases in these posts. The results are already translated to English and attached in the appendix. I also plot a word cloud for the leading words and phrases in Figure 3.1.

As we can see from the word cloud, in Guba, most posts are about the price

Figure 3.1: Word Cloud of Users' Posts

change of the stocks. Users also care about whether "main funds", which usually refers to the large investment funds, brokerage, and insurance companies, will buy in or sell out this stock. Another interesting intuition from this word cloud is that users are usually pessimistic, given that "Drop", "Junk", "Leave", "Run", etc. negative words take a large part of the words. This indicates that individual users likely tend to post negative comments, and therefore, their sentiment index should be lower.

3.3 Sentiment Analysis of Textual Data

So far, I have chosen the target social media platform and constructed the stock portfolio. By inputting the codes of stocks in the stock portfolio, the web crawler described in Part 3.1 will automatically gather the posts in our target forum. This gives us a large table with read times, comment times, posting contents and time for each chosen stock. However, textual data cannot be directly put into

our model. As mentioned before, to establish a proxy for investor sentiment, I need to utilize the natural language processing (NLP) and sentiment analysis technique.

In Part 2.2, I briefly reviewed the development history of NLP and sentiment analysis techniques. Through this review, I also justified the reliability of these techniques. In economics and financial research field, sentiment analysis technique is not a novelty. Hitherto, there are several English dictionaries even specially designed for English financial texts. Some of them are totally open sourced, such as General Inquirer (Hartman et al., 1967), Harvard IV Dictionary and Loughran and McDonald's Accounting and Finance Dictionary (Loughran and McDonald, 2013). From previous work, we know that Chinese sentiment analysis is a little more complicate since Chinese sentences need to be segmented to words to apply lexicon-method. There are also frequently used Chinese financial dictionaries, such as Bian et al.'s Chinese Financial Sentiment Dictionary (CFSD) by (Bian et al., 2019).

Like English sentiment analysis, lexicon-based method is relatively conventional in pervious literature. Currently, supervised machine-learning methods are also frequently used by Chinese researchers, such as the normal support vector machine, and deep learning convolutional neural network. Although J. Li et al. have used deep learning convolutional neural network to analysis the Chinese stock market sentiment, their work is not open-source (Li et al., 2019). Actually, it is very hard to find reliable open-source packages to analysis Chinese financial texts. To reimburse this problem, this thesis focuses on open-source pre-training algorithm which can be fine-tuned to make the algorithm more adapted to financial textual data. One of such leading up-to-date algo-

rithms is the Senta-BiLSTM model of SKEP from a research team from Baidu Inc., and University of Science and Technology of China (Tian et al., 2020).

The Senta-BiLSTM is a part of Baidu's Sentiment Knowledge Enhanced Pre-training (SKEP) model and it is totally opened to public and easy to access. Based on PArallel Distributed Deep LEarning (PaddlePaddle), an efficient, flexible and scalable deep learning platform developed by Baidu engineers comparable to TensorFlow or PyTorch, SKEP has a better performance than other similar models such as RoBERTa (Liu et al., 2019) in the tasks of Sentence-level Sentiment Classification, Aspect-level Sentiment Classification and Opinion Role Labeling (Tian et al., 2020). The structure of SKEP is composed of two main parts. The first part is called "sentiment masking". In this part, the input sentences are initially segmented, and the sentiment knowledge is automatically mined relying on an unsupervised statistical method just like Turney's (Turney, 2002). The sentiment knowledge includes sentiment words, such as "fast", "appreciate", etc. It also includes sentiment polarity, which is whether a word is positive and negative; and aspect-sentiment pairs, which are the mention of an aspect and its corresponding sentiment word. Then the next step in this part is "masking". To be short, the "masking" is to corrupt the input sentences by masking the sentiment information. It includes traditional words and phrases, and it also includes aspect-sentiment pairs unlike other classical pre-training models. The second part of SKEP model is "sentiment pre-training objectives". Corresponding to the sentiment knowledge mining, the pre-training objectives are to recover the corrupted sentences, including the sentiment words, polarity, and the aspect-sentiment pairs. Thus, through these two parts, the model can be trained with the "sentiment pre-training objectives" which are supervised by recovering sentiment information. Hence, the model can successfully complete

its task by self-supervised training.

As a pre-training model, it can be fine-tuned to make it work better on specific dataset. The authors tested SKEP on various datasets of different kinds of tasks. For instance, on the task "Sentence-level Sentiment Classification", which is exactly what we need to make SKEP work on, SKEP shows a very high accuracy (ACC) on Chinese language dataset, such as ChnSentiCorp and NLPCC2014-SC. The ACC are respectively 96.50% and 83.53%.

Senta-BiLSTM is an openly access version of SKEP. It can be easily deployed though Baidu's PaddleHub and simple Python commands. The speed of Senta-BiLSTM is decent. My test shows that without fine-tuning, to analyze 20 long sentences, it will take only 2 seconds. The output of Senta-BiLSTM are three variables: the positive and negative probability of the sentence and the sentiment prediction – whether it is positive, negative, or neutral.

My job in the model is "fine-tuning". In other words, by inputting my own dataset and then training the model, I can achieve a better performance on financial texts through these small adjustments. This process can enhance the reliability of my model.

My sentiment index for each posting is constructed as the following formula:

$$SSI = PP - NP \tag{3.1}$$

And SSI is the the sentiment index for this post/sentence, PP is the positive probability and NP is the negative probability.

Usually there are hundreds of posting each day for a single stock, so a daily

sentiment index might be more useful for us:

$$SI_t = \sum_i SSI_{i,t} \frac{1}{2} \left(\frac{RT_{i,t}}{\sum_i RT_{i,t}} + \frac{CT_{i,t}}{\sum_i CT_{i,t}} \right) \tag{3.2}$$

In this formula, SI_t is the daily sentiment index at time t for a specific stock; $SSI_{i,t}$ is the sentiment index for posting i at time t; $RT_{i,t}$ is the read times for posting i at time t and $CT_{i,t}$ is the comment times for posting i at time t. To be short, here I define the daily sentiment index as the sum of each posts weighted by their read times and comment times. Here, this formula puts more weight on the comment times. It is because comments can better indicate whether a post is more valued by other users.

Before calculating SI_t, an important task is to distinguish individual users' posts from institutional user's posts. After segmenting individual and institutional users' posts, I can calculate individual investors' daily sentiment index using formula 3.2. Here, I denote the daily sentiment index for individual users as SIU_t and the daily sentiment index for institutional users as SIP_t.

After filtering out institutional users' posts here, I adapt the data from Uqer to calculate institutional investors' sentiment. The previous part states that the variables which I can directly get from Uqer are the stock code, the published date of news and report, the score of relation and the sentiment score of the news/report. The score of relation $SR_{i,t}$ measures how much the news or analysis report is related to the stock/company, where 1 means they are directly related and 0 means they are not related. The sentiment score $SSN_{i,t}$ is similar to my constructed sentiment score for individual users' post, where 1 means the news is totally positive and -1 means the news is totally negative.

Weighted by the score of relation, I use the similar formula to calculate the

daily data of institutional investor's sentiment:

$$SIP_t = \sum_i SSN_{i,t} \left(\frac{SR_{i,t}}{\sum_i SR_{i,t}} \right) \tag{3.3}$$

Now I have successfully transformed the textual data to numerical sentiment indices. I also get individual and institutional investors sentiment index. In next part, I will introduce the data chosen for Chinese stock market and how this thesis acquires these data.

3.4 Data for Stock Markets

The last step is to acquire related data for Chinese stock market. In Part 2.1, I reviewed previous work on the relationship of investor sentiment and stock market. As we can see, large amounts of papers pay attention to stock returns, including the work of Fisher and Statman, Brown and Cliff, Schmeling, etc. (Fisher and Statman, 2000; Brown and Cliff, 2005; Schmeling, 2009). Other researchers also care about trading volumes, trading volatility and the activity of mutual funds besides the gains from stock market, such as the work of Das and Chen, Joseph et al. and Da et al. (Da et al., 2015; Das and Chen, 2007; Joseph et al., 2011). These papers suggest the correlation of sentiment and all aspects of stock market.

In line with previous experience, in this thesis I focus on the four variables of stock market. The initial one is the closing price, and the derived price change, stock returns, etc. I also care about the trading volume and trading amount, to investigate whether the sentiment will impact the market activity. Last but not least, I care about the turnover rate. It can help in investigating whether the sentiment will influence the fluidity of the stock.

The basic data can all be acquired from the official site of Shenzhen Stock Exchange (SZSE)[2] and other financial websites. To improve the efficiency, this thesis uses an open-source package Akshare[3] to gather the related data. This package calls the API of Sina Finance and Tencent Finance to automatically collect the shares data. This package also provides other functions to call more advanced data.

3.5 Descriptive Statistics

This part briefly introduces how the variables in the thesis are constructed. Then, this part provides descriptive statistics for the basic data and derived variables. As aforementioned, the data this thesis uses can be divided into two major parts: the sentiment-related data and the stock market related data. The sentiment-related data includes the proxies for individual and institutional investors sentiment. The stock market related data includes the closing price, the trading amount and volume and the turnover rate. By using closing prices, I also calculate the stock return and the daily volatility.

The daily sentiment indices for institutional investors (SIP_t) and individual investors (SIU_t) are calculated by formula 3.2 and 3.3. When the indices are closing to 1, it indicates that the sentiment on this day is more optimistic; when the indices are closing to -1, it indicates that the sentiment on this day is more pessimistic.

The daily stock market data includes the closing price CP_t, the trading

[2]Official Website: http://www.szse.cn/English/
[3]Their project: https://github.com/jindaxiang/akshare

amount TA_t and volume TV_t and the turnover rate TR_t. All of them can be directly accessed. The relationships of the variables can be described using the following formula.

$$TA_t = CP_t TV_t \tag{3.4}$$

$$TR_t = \frac{TV_t}{OS_t} \tag{3.5}$$

In formula 3.5, OS_t denotes outstanding shares. From the closing price CP_t, we can calculate the stock returns SR_t and volatility VOL_t by using the exponentially weighted moving average (EWMA) formula.

$$SR_t = \ln \frac{CP_t}{CP_{t-1}} \tag{3.6}$$

$$VOL_t^2 = \lambda VOL_{t-1}^2 + (1-\lambda) SR_{t-1}^2 \tag{3.7}$$

Here, I choose $\lambda = 0.94$ in line with common approaches in RiskMetrics. This indicates a slower decay in the time series. The EMWA method improves the simple standard deviation by setting up different weight in periodic returns.

However, these variables of sentiment and stock market above are for each single stock and there are 18 of them. To better investigate the overall and cross-section effect of sentiment, we will also need synthesized data. Recall the definition and composition of market indices, a good way to synthesize the data is to use their market value as weights. For the sentiment indices, our synthesized data is constructed by the following formula.

$$ASIP_t = \sum_j SIP_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{3.8}$$

$$ASIU_t = \sum_j SIU_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{3.9}$$

In these formulae, $ASIP_t$ and $ASIU_t$ denote the synthesized sentiment indices at time t. $SIP_{t,j}$ and $SIU_{t,j}$ denote the sentiment indices for stock j at time t. And $MV_{t,j}$ denotes the market value of stock j at time t.

Table 3.2: Descriptive Statistics of Synthesized Sentiment Indices (Before/After IPO)

	IPO	Min	Max	Mean	Std. Dev.	Kurtosis	Skewness
$ASIP_t$	Before	-0.524	0.475	0.147	0.157	7.164	1.364
$ASIP_t$	After	-0.110	0.517	0.179	0.101	0.213	0.100
$ASIU_t$	Before	-0.484	0.738	-0.058	0.170	7.051	1.112
$ASIU_t$	After	-0.557	0.239	-0.250	0.133	2.714	1.351

Similarly, I can use these formulae to synthesize the market data. Choose the date of IPO, i.e., August 24th, 2020, as baseline and set the base index for the stock portfolio as 100, we can have the formulae for the stock market data.

$$AI_t = \frac{\sum_j MV_{t,j}}{\sum_j MV_{0,j}} 100 \tag{3.10}$$

$$ATA_t = \sum_j TA_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{3.11}$$

$$ATV_t = \sum_j TV_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{3.12}$$

$$ATR_t = \sum_j TR_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{3.13}$$

$$ASR_t = \ln \frac{AI_t}{AI_{t-1}} \tag{3.14}$$

$$AVOL_t = \lambda AVOL_{t-1}^2 + (1-\lambda) ASR_{t-1}^2 \tag{3.15}$$

The notations are the same as sentiment data. $MV_{0,j}$ is the market value of stock j at the baseline date: i.e., August 24th, 2020. The following tables 3.2 and 3.3 show the descriptive statistics for the synthesized data.

Table 3.3: Descriptive Statistics for Synthesized Stock Market Data (After IPO)

	Min	Max	Mean	Std. Dev.	Kurtosis	Skewness
AI_t	68.53	117.56	91.73	7.67	0.27	-0.25
ASR_t	-0.11	0.16	0.00	0.03	4.58	0.33
$AVOL_t$	0.00	0.06	0.03	0.01	0.57	0.38
ATV_t	1628996.07	21656992.88	4753957.81	3256189.77	7.88	2.36
ATA_t	179793634.43	2324493109.73	503639517.70	322136553.71	8.49	2.40
ATR_t	4.63	61.72	13.56	9.44	8.01	2.39

As we can see from the tables, an interesting finding in line with previous part is that individual investors are usually over pessimistic. Before IPO, individual investors' mood is nearly neutral while large number of posts sentiment scores are 0. The reason might be that before IPO, investors usually discuss impartial matters on the forum. After IPO, when the shares start to trade, investors begin to discuss the stocks themselves. And the table shows an overwhelming pessimistic mood. The average sentiment score is only -0.25, and standard deviance shows that major part of data concentrate around this average score. For institutional investors, the trend is exactly the different. Before IPO, the average sentiment score is 0.147. After IPO, it goes up to 0.179 and standard deviance decreases. It implies an increasing confidence of professionals.

CHAPTER 4

METHODS

4.1 Theoretical Analysis

Many classical behavior economics and finance theories imply a possible correlation of stock market and investor sentiment. As reviewed in Part 2.1, one of the classical model is the DSSW noise trader risk model (De Long et al., 1990). The DSSW model has two types of agents: the sophisticated one in ratio $1-\mu$ and the noise trader in ratio μ who does not have rational expectations. The false belief of noise trader is an i.i.d. normal random variable $\rho_t \sim N\left(\rho^*, \sigma_p^2\right)$. In time $t+1$, agents try to maximize their utility $U = -e^{-(2\gamma)w}$, where γ is the absolute risk aversion, w is the wealth gained at $t+1$. When the return of holding a risky asset is normally-distributed, then maximizing U is equivalent to maximizing $\bar{w} - \gamma\sigma_w^2$, where $\bar{w}$ is the expected gain and σ_w^2 is the variance of asset one period ahead.

Then, the DSSW model assumes sophisticated agent holds λ_t^i amount of risky asset, and noise trader holds λ_t^n. By writing λ_t^i and λ_t^n as the function of P_t, $E_t(P_{t+1})$ and $E_t\left(\sigma_{P_{t+1}}^2\right)$ and using equilibrium condition $\mu\lambda_t^n + (1-\mu)\lambda_t^i = 1$, the model derives the formula of the asset price at time t (De Long et al., 1990).

$$P_t = 1 + \frac{\mu(\rho_t - \rho^*)}{1+r} + \frac{\mu\rho^*}{r} - 2\frac{\gamma\mu^2\sigma_p^2}{r(1+r)^2} \tag{4.1}$$

In this equation, r denotes the interest rate. From equation 4.1 we can see that the "false belief" of noise traders can indeed impact the asset price. The term $\frac{\mu(\rho_t-\rho^*)}{1+r}$ actually captures the fluctuation of the asset price because of the misperceptions of noise traders. When ρ_t is higher than ρ^*, noise traders are over

optimistic, and they will push the asset price higher. When ρ_t is lower than ρ^*, noise traders are over pessimistic, and they will push the asset price lower. When μ is higher, the volatility term will also go higher. It implies that the noise traders' false belief will impact the asset price changes and volatility, and more noise traders in the market will enhance the fluctuation of the asset price. The term $\frac{\mu\rho^*}{r}$ captures the effect to asset price when ρ^*, noise traders' misperceptions, is not zero. When noise traders are averagely bullish, then the price will be higher than it would be; when noise traders are averagely bearish, then the price will be lower than it would be. The authors called this effect "price pressure". The term $2\frac{\gamma\mu^2\sigma_p^2}{r(1+r)^2}$ denotes the compensation since there exists risk that noise traders might become bearish, and the price will fall because of that. This risk only comes from traders' false belief, and have no relation to the asset itself and macroeconomics environment (De Long et al., 1990).

In fact, investors sentiment is a kind of "false belief" (Barberis et al., 1998; Baker and Wurgler, 2006). The DSSW model implies that noise traders will significantly influence the stock market, and investors sentiment can impact the asset price, the returns, and the price volatility. The model also argues that since noise traders bear an unfair amount of risk which they own created, noise traders tend to have a higher expected return than sophisticated investors.

Sentiment is also a core concept in the behavior approaches to asset pricing. The model of Shefrin and Belotti shows that the log stochastic discount factor (log-SDF) is the sum of a sentiment process and another stochastic process based on aggregate consumption growth. This shows the asset prices are only efficient when the sentiment Λ is uniformly zero, indicating that there are neither aggregate belief distortion nor SDF aggregation bias. Their model also

solves for the sentiment premium. The result shows that the expected return of any risky asset should be the sum of three components: the interest rate, the risk premium, and the sentiment premium. This sentiment premium reflects sentiment-based risk and captures the mispricing in respect to the risk-free rate and the price dynamics (Shefrin and Belotti, 2008).

Overall, most behavior economists theories believe that sentiment is a synonym of "error". According to the formal definition of Shefrin and Belotti, "sentiment" in behavior finance is the "aggregate errors of investors being manifest in security prices" (Shefrin and Belotti, 2008). From the DSSW model and the sentiment premium, we can tell that the influence of investor sentiment on stock market is supported by current theories.

4.2 Quantitative Methods

This part, I examine the relationship of investor sentiment and stock market by using various different quantitative methods. First I focus on the after-IPO data, from the IPO date August 24th, 2020 to January 26th, 2021. Then I turn our sight on the pre-IPO data, which is before the IPO date August 24th, 2020.

According to the previous theory analysis, we know that theories such as DSSW model predicts that a higher sentiment will push the asset price higher (De Long et al., 1990). Consider an arbitrage situation, where there exists another asset that is not influenced by the "noise trader", or in my words, the investors sentiment. Then arbitragers will sell out the asset that has been pushed to a higher price by optimistic sentiment and buy in the other asset. If we view the sentiment as a sudden impulse, this means that while the asset price will be

pushed higher temporally, the price will fall down after this short time. For a low sentiment, vice versa, the asset price will be pushed lower temporally, the price will rise after. This prediction is known as the "return reversals" (Da et al., 2015).

4.2.1 Exploratory OLS of Sentiment Indices and Stock Market Variables

Initially, here I use a simple linear regression model to find evidence of the possible relationship of investor sentiment and stock market variables. In fact, this is a frequently-used method in the research of investor sentiment and stock market, such as in the paper of Da et al., where their similar model found evidence of the association of their sentiment proxy and stock returns (Da et al., 2015).

I define the OLS model as follows:

For stock prices:

$$AI_{t+k} = \beta_0 + \beta_1 ASIP_t + \beta_2 ASIU_t + \sum_m \gamma_m CTR_t^m + \mu_{t+k} \tag{4.2}$$

In formula 4.2, AI_{t+k} is the broad equity index for our stock portfolio at time t+k where $k \geq 0$. CTR_t^m denotes the control variables. Like classical theories, it includes the lagged equity index AI up to three terms, the Shenzhen Stock Exchange (SZSE) whole index (denoted as $SZSEI_t$), the interest rate (Here, I use the Shanghai Interbank Offer Rate (O/N), abbr. Shibor O/N, as the interest rate[1]. It is denoted as IR_t.), the volatility index of the stock portfolio (denoted as $AVOL_t$ as aforementioned.). Many literature mentioned the link between politics

[1]Official Website: http://www.shibor.org/shibor/web/html/index_e.html

and financial markets (Addoum and Kumar, 2016), therefore I also include the change of economic policy uncertainty index (denoted as ΔEPU_t) in our model. μ_{t+k} denotes the error term. Hence, similar to formula (4.2), I can define other regressions.

For stock returns:

$$ASR_{t+k} = \beta_0 + \beta_1 ASIP_t + \beta_2 ASIU_t + \sum_m \gamma_m CTR_t^m + \mu_{t+k} \tag{4.3}$$

For fund flows data, i.e., trading volume, trading amount and turnover rate, we have:

$$ATA_{t+k} = \beta_0 + \beta_1 ASIP_t + \beta_2 ASIU_t + \sum_m \gamma_m CTR_t^m + \mu_{t+k} \tag{4.4}$$

$$ATV_{t+k} = \beta_0 + \beta_1 ASIP_t + \beta_2 ASIU_t + \sum_m \gamma_m CTR_t^m + \mu_{t+k} \tag{4.5}$$

$$ATR_{t+k} = \beta_0 + \beta_1 ASIP_t + \beta_2 ASIU_t + \sum_m \gamma_m CTR_t^m + \mu_{t+k} \tag{4.6}$$

To summarize, by interpreting the β_0,β_1,β_2 and their significance, I can find possible evidence of the impact of sentiment in stock market. I can also find how the sentiment impacts the stock market. Meanwhile, by adjusting the k, I can explore the potential lagged effect and the so-called "return-reversal" phenomenon.

4.2.2 VAR Model and Impulse-Response Analysis

Impulse-response is an effective method to investigate the dynamic reflection of a system generated from an inside or outside impulse. Here, I use the VAR model and its related impulse-response analysis method to explore the impulse generated from the sentiment indices.

Compared to the linear regression models and models such as ARIMA model, VAR model takes the lagged value of other variables into account. A typical $VAR(p)$ model can be written as the following form:

$$r_t = \phi_0 + \sum_{i=1}^{p} \Phi_i r_{t-i} + a_t \tag{4.7}$$

In this formula 4.7, r_t is the k-multivariate time series; ϕ_0 is the vector of the constant term; Φ_i is a k-order square matrix; and a_t is the weakly stationary error term with $E(a_t) = 0$ and $var\,(a_t) = \mathbf{\Sigma} > 0$. Here, I assume that a_t is normally distributed, and $a_t = (a_{1t}, a_{2t}, \ldots, a_{kt})^T$.

Formula 4.7 can be also written is a shorten form:

$$P\,(B)\,r_t = \phi_0 + a_t \tag{4.8}$$

And $P\,(B) = I - \sum_{i=1}^{p} \Phi_i B^i$. Transform formula 4.7 into an infinite MA solution function, I have:

$$r_t = \mu + \sum_{i=0}^{\infty} \Psi_i a_{(t-i)} \tag{4.9}$$

This formula shows that Ψ_i is the impact coefficients of past information to time series r_t. Then the elements of Ψ_i can be regarded as the coefficients of the Pulse Response Function of r_t. If define:

$$\underline{\Psi}_n = \sum_{i=0}^{n} \Psi_i \tag{4.10}$$

Then $\underline{\Psi}_n$ in formula 4.10 is known as the cumulative impulse.

Here, I use the VAR model and impulse-response function to testify the following dependent variables: AI_t, ASR_t, ATV_t and $AVOL_t$.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Results of Exploratory OLS Models

Using the time series data from the IPO date August 24th, 2020 to January 27th, 2021, here I report parts of the regression results of model 4.2 and model 4.5. Here I set the dependent variables up to four ensuing trading days. However, in most of the cases, coefficients become insignificant from the second ensuing day. Therefore, I only report the regression results to the second ensuing day.

(1) Model 4.2: price and sentiment in Table 5.1.

(2) Model 4.5: trading volume and sentiment in Table 5.2.

From the regression results of model 4.2 and model 4.3, an obvious finding is that individual investors' sentiment can significantly impact the stock price and stock return. This kind of impact can be both contemporary and ensuring. For example, in Table 5.1, as we can see from the first column, when the sentiment index of individual investors increases by one, the broad equity index can increase by 9.57. From the second column, we can notice that the impact increases to a higher level in the following trading day: when the sentiment index of individual investors increases by one, the broad equity index can increase by 11.37. This finding is reasonable given that traders make trading decision after viewing these online posts. For the second ensuing day, the impact of individual sentiment becomes insignificant . This result is also reasonable if considering the sentiment index is a short-time impulse. Surprisingly, the OLS models cannot provide proofs that individual investor's sentiment can impact

Table 5.1: OLS Result of Model 4.2

	(1) $AI(t)$	(2) $AI(t+1)$	(3) $AI(t+2)$
Constant	55.598***	99.807***	139.243***
	(11.53)	(13.65)	(16.07)
$ASIU$	9.57**	11.37**	3.230
	(3.91)	(4.642)	(5.45)
$ASIP$	1.570	1.570	0.287
	(2.87)	(3.45)	(4.06)
$AVOL$	-133.72***	-253.14***	-339.84***
	(46.09)	(54.81)	(64.77)
$SZSEI$	-0.0016***	-0.0032***	-0.0049***
	(0.001)	(0.001)	(0.001)
ΔEPU	0.0098***	0.0155***	0.0234***
	(0.003)	(0.004)	(0.005)
IR	70.429	164.774**	163.289*
	(55.07)	(67.26)	(81.58)
$AI(t-1)$	0.7733***	0.7292***	0.4303***
	(0.099)	(0.117)	(0.14)
$AI(t-2)$	0.003	-0.321**	-0.095
	(0.11)	(0.13)	(0.146)
$AI(t-3)$	-0.074	0.109	-0.004
	(0.07)	(0.09)	(0.10)
Adj. R^2	0.881	0.835	0.773

This table presents OLS regression result of relation of individual and institutional investor's sentiment with stock price. I also relate the stock price in the future two days to $ASIU$ and $ASIP$. The control variables are the lagged values of AI up to three days, the volatility AVOL, the total index of Shenzhen Exchange $SZSEI$, the change of economic policy uncertainty index ΔEPU and the interest rate IR. In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

the stock price and stock return. I will reexamine this using VAR models. From the control variables, we can tell the volatility, Shenzhen Exchange Index, the change of economic policy uncertainty index and the lagged stock price/return have a significant impact on the stock price/return. These results are similar to the results of the control variables of Da et al.'s work (Da et al., 2015).

From the regression results of model 4.5 and model 4.6, I cannot find evi-

Table 5.2: OLS Result of Model 4.5

	(1)	(2)	(3)
	$ATV(t)$	$ATV(t+1)$	$ATV(t+2)$
Constant	-1.226e+07*	9.061e+06	2.275e+07***
	(6.38e+06)	(6.83e+06)	(6.63e+06)
$ASIU$	1.767e+06	3.149e+06	2.519e+06
	(2.17e+06)	(2.32e+06)	(2.25e+06)
$ASIP$	8.346e+05	-1.12e+06	1.069e+06
	(1.59e+06)	(1.72e+06)	(1.67e+06)
$AVOL$	2.877e+08***	2.107e+08***	1.53e+08***
	(2.55e+07)	(2.74e+07)	(2.67e+07)
$SZSEI$	276.1721	-480.4728	-1009.058***
	(297.638)	(319.386)	(311.314)
ΔEPU	762.6221	2879.0961	3788.12**
	(1787.296)	(1914.301)	(1859.776)
IR	-7.608e+07	-3.743e+07	-6.013e+06
	(3.05e+07)	(3.37e+07)	(3.36e+07)
$AI(t-1)$	1.504e+05***	5.841e+04	-4.558e+04
	(5.49e+04)	(5.86e+04)	(5.67e+04)
$AI(t-2)$	5.231e+04	2.74e+04	3.601e+04
	(5.83e+04)	(6.23e+04)	(6.04e+04)
$AI(t-3)$	-1.426e+05***	-1.15e+05***	-8.135e+04*
	(4e+04)	(4.27e+04)	(4.14e+04)
Adj. R^2	0.706	0.626	0.620

This table presents OLS regression result of relation of individual and institutional investor's sentiment with trading volume. I also relate the stock price in the future two days to $ASIU$ and $ASIP$. The control variables are the lagged values of AI up to three days, the volatility AVOL, the total index of Shenzhen Exchange $SZSEI$, the change of economic policy uncertainty index ΔEPU and the interest rate IR. In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

dence that the individual or institutional investors' sentiment can impact the stock market's activity: neither they will significantly influence the trading volume and the turnover rate. For the ensuing days, the significance of individual and institutional investors' sentiment slightly increases, indicating a possible better predictive meaning in the future. This finding will be testified in the VAR models. The regression results of the control variables also cannot provide evidence that the volatility, Shenzhen Exchange Index, and the change of economic

policy uncertainty have influence on the trading volume and turnover rate.

We also cannot observe a so-called "return-reversal" phenomenon in these OLS models. This result seems to be reasonable considering the source of my sentiment indices. As I mentioned in previous part, my daily sentiment indices are generated from real-time forum posts. Real-time means that there will be hundreds of posts for each single stock every day. Consider an impactful user (usually with a high read times or comment times) posts something positive in the forum at 10:00 AM, like "The stock price will definitely rise up in the near future.". Then if some users or investors read and believe his opinion, they will start to buy in this stock if they have enough funds to invest. Or when the overall sentiment in the forum is pessimistic, those who believe ideas in the forum will tend to sell their stocks. These trading will make the stock price rise or fall. In fact, there has been evidence from a dedicated financial experiment that some investors indeed trade relying on the Internet posts (Ammann and Schaub, 2020). If their actions regarding these posts are immediate and decisive, then the impact to the stock price or other related variables can be viewed as instantaneous in a daily base. Due to some theories, the "reversal" can be caused by a liquidity shock. However, we cannot observe a correlation between the sentiment and trading volumes, therefore it might do not exist such a liquidity shock. This might be the reason why I cannot observe a short-term "reversal" in my OLS models.

Table 5.3: ADF Results of Main Variables

	ADF	p-Val.	Lagged	1%	AIC	Pass
AI	-2.45	0.13	0	-3.49	423.82	FALSE
ΔAI	-7.62	0.00	2	-3.50	426.26	TRUE
ASR	-7.00	0.00	2	-3.50	-392.49	TRUE
ATV	-1.35	0.60	12	-3.50	2818.73	FALSE
ΔATV	-5.91	0.00	11	-3.50	2787.60	TRUE
$AVOL$	-1.85	0.35	5	-3.50	-907.05	FALSE
$\Delta AVOL$	-6.45	0.00	4	-3.50	-894.46	TRUE
$ASIU$	-2.70	0.07	11	-3.50	-243.76	FALSE
$\Delta ASIU$	-4.50	0.00	12	-3.50	-231.78	TRUE
$ASIP$	-8.47	0.00	0	-3.49	-176.07	TRUE
$SZSEI$	-0.38	0.91	0	-3.49	1211.81	FALSE
$\Delta SZSEI$	-9.83	0.00	0	-3.49	1199.25	TRUE
IR	-1.87	0.35	2	-3.50	-821.10	FALSE
ΔIR	-9.42	0.00	1	-3.50	-810.55	TRUE

This table presents ADF result of main independent and dependent variables. In this table, Δ denotes the first-order difference transformation. I only consider 1% significance in this table. If the p-value is higher than 0.05, I still consider that it fails the ADF test.

5.2 Results of VAR Models

A huge defect in the previous OLS models is the spurious regression or the spurious relationship. This kind of phenomenon is due to the non-stationary time series. The spurious regression problem can lead to a confusing evidence that the independent and dependent variables are correlated in a linear relationship. And it can damage the reliability of our OLS models results. However, the VAR model can fix this problem.

To handle the spurious regression problem, one important step before constructing the VAR model is to perform the Unit Root Test. The goal of this test is to examine the stationary of the time series. One frequently used Unit Root Test method is ADF, or the Augmented Dickey–Fuller test. Here, I report the results of ADF tests in the following Table 5.3.

Table 5.4: Cointegration Test Results

	Coint	p-Val	1%	Pass
$AI - ASIU$	-3.28	0.06	-4.00	FALSE
$AI - SZSEI$	-2.45	0.30	-4.00	FALSE
$AI - IR$	-1.72	0.67	-4.00	FALSE
$ATV - ASIU$	-1.07	0.89	-4.00	FALSE
$ATV - SZSEI$	-2.37	0.34	-4.00	FALSE
$ATV - IR$	-2.51	0.27	-4.00	FALSE
$AVOL - ASIU$	-1.73	0.66	-4.00	FALSE
$AVOL - SZSEI$	-4.30	0.00	-4.00	TRUE
$AVOL - IR$	-1.36	0.81	-4.00	FALSE

This table presents cointegration test result of main independent and dependent variables. I only consider 1% significance in this table. If the p-value is higher than 0.01, I still consider that it fails the cointegration test.

As we can see from Table 5.3, most of the variables expect the sentiment of institutional investors are not stationary. Instead, I will need to conduct the first-order difference transformation to make the variable stationary. Here, I use Δ to notate the first-order difference transformation. The ADF test of these variables after the first-order difference transformation shows that they are stationary, and therefore do not need a second-order difference transformation.

The second part is the cointegration test. The cointegration test is designed for the non-stationary times series and is to examine the possible presence of cointegration between the non-stationary times series. The following Table 5.4 presents the results of the cointegration test of the independent and dependent variables.

As we can see from Table 5.4, commonly I cannot use the origin time series since most of the variable's combinations fail the cointegration test except the volatility $AVOL$ and the Shenzhen Exchange Index $SZSEI$. This indicates that these variable pairs do not exist a long-term equilibrium relationship. Therefore, I should use the first-order difference time series variables to construct the VAR

Table 5.5: Lag Order Selection

	Lag	AIC	BIC	FPE	HQIC
Model VAR-AI	0	-9.17	-9.037*	1e-4	-9.118*
	1	-9.28	-8.47	9.31e-5	-8.95
	2	-9.18	-7.68	1e-4	-8.57
	3	-9.37*	-7.19	8.67e-5*	-8.49
Model VAR-ASR	0	-18.17	-18.03*	1.28e-8	-18.11*
	1	-18.28	-17.46	1.15e-8	-17.95
	2	-18.16	-16.66	1.31e-8	-17.55
	3	-18.33*	-16.15	1.11e-8*	-17.45
Model VAR-ATV	0	17.10	17.23*	2.66e+7	17.15*
	1	17.07	17.88	2.58e+7	17.40
	2	17.04	18.54	2.53e+7	17.64
	3	16.83*	19.00	2.06e+7*	17.70
Model VAR-AVOL	0	-23.49*	-23.35*	6.30e-11*	-23.43*
	1	-23.45	-22.63	6.54e-11	-23.12
	2	-23.35	-21.85	7.27e-11	-22.75

This table presents the selection criteria result of the three models. * highlights the minimums.

model.

One last step before constructing the VAR model is to select the VAR lag order. The statistical theories provide several difference selection criteria such as AIC, BIC, FPE and HQIC. Here, I report the results of these criteria in the following Table 5.5.

Based on the AIC and FPE criteria, the VAR lag order of model VAR-AI, model VAR-ASR and model VAR-ATV is chosen to be three. I also choose VAR lag order of model VAR-AVOL as zero since all the four criteria indicate zero lag order is the best. Therefore, here I can discard the VAR-AVOL model in the following analysis. After setting the VAR lag order, I can fit the three models. After fitting the models, I can plot the autocorrelation plots of residuals of the three models. Here, the autocorrelation plot of Model VAR-AI is shown in below Figure 5.1.

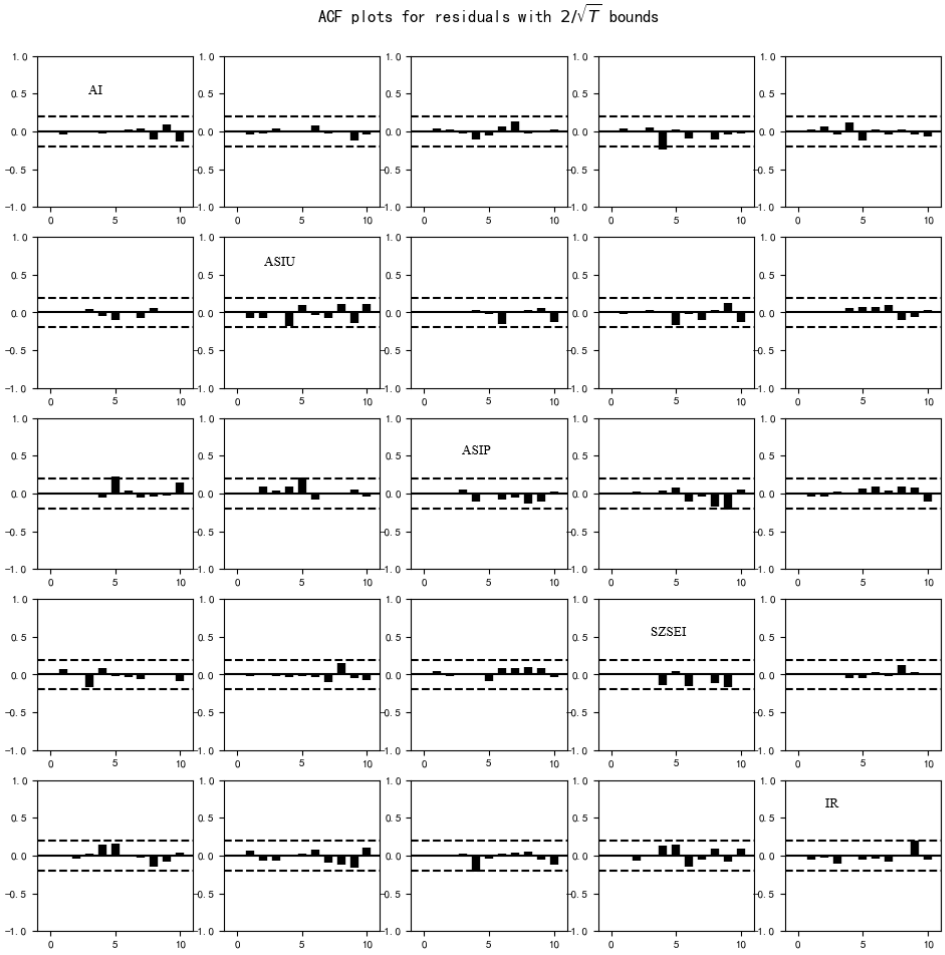

Figure 5.1: ACF Plots for Residuals - Model VAR-AI

From the above figure and figures in the appendix, for all the three models, basically all of the five variables' residuals are within the $\pm\frac{2}{\sqrt{T}}$ border. This indicates that there is no obvious autocorrelation of the residuals . Therefore, we can believe that the models are decent approximations of the data. These results further validate the VAR models.

For the VAR models, Granger test can help to examine whether a variable's lagged value $x(t)$ can be introduced to the model of $y(t)$ as an endogenous variable and explain the variations of $y(t)$. Here, I only care about whether the sen-

Table 5.6: Granger Test Result of Model VAR-AI, Model VAR-ASR and VAR-ATV

H_0	χ^2	p-Val	Reject
$\Delta ASIU$ does not Granger-cause ΔAI	5.61	0.13	FALSE
$ASIP$ does not Granger-cause ΔAI	1.29	0.73	FALSE
$\Delta ASIU$ does not Granger-cause ASR	6.28	0.09*	TRUE
ASR does not Granger-cause $\Delta ASIU$	0.85	0.84	FALSE
$ASIP$ does not Granger-cause ΔASR	1.60	0.66	FALSE
$\Delta ASIU$ does not Granger-cause Δ ATV	3.21	0.36	FALSE
$ASIP$ does not Granger-cause ΔATV	12.31	0.006***	TRUE
ΔATV does not Granger-cause $ASIP$	1.99	0.58	FALSE

In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

timent of individual and institutional investors can Granger-cause the variation of the stock price AI, the return ASR and the trading volume ATV. Due to the lagged order in model VAR-AVOL is 0, I cannot perform Granger test in this model. Using Wald χ^2 statistics, the Granger test result is reported below in Table 5.6.

The results of the Granger test show that neither the change of individual investors' sentiment nor institutional investors' sentiment can help predict the stock price. Although $\Delta ASIU$ has a higher χ^2 than $ASIP$, it still fails the test even in 10% level of significance. For the change of stock return ΔASR, the change in individual investors sentiment has some help in predicting its variance. And in turn, the stock return cannot help explaining the variance of the change of individual sentiment. Therefore, we can believe that the change in individual sentiment is an essential variable in predicting the stock return. For the trading volume, I cannot find proofs that the change of individual investors' sentiment can help explain the change of trading volume. But the institutional investors' sentiment level can indeed help predict the change of trading volume significantly. Meanwhile, because the change of trading volume cannot help predicting the institutional investors' sentiment level, we can believe the

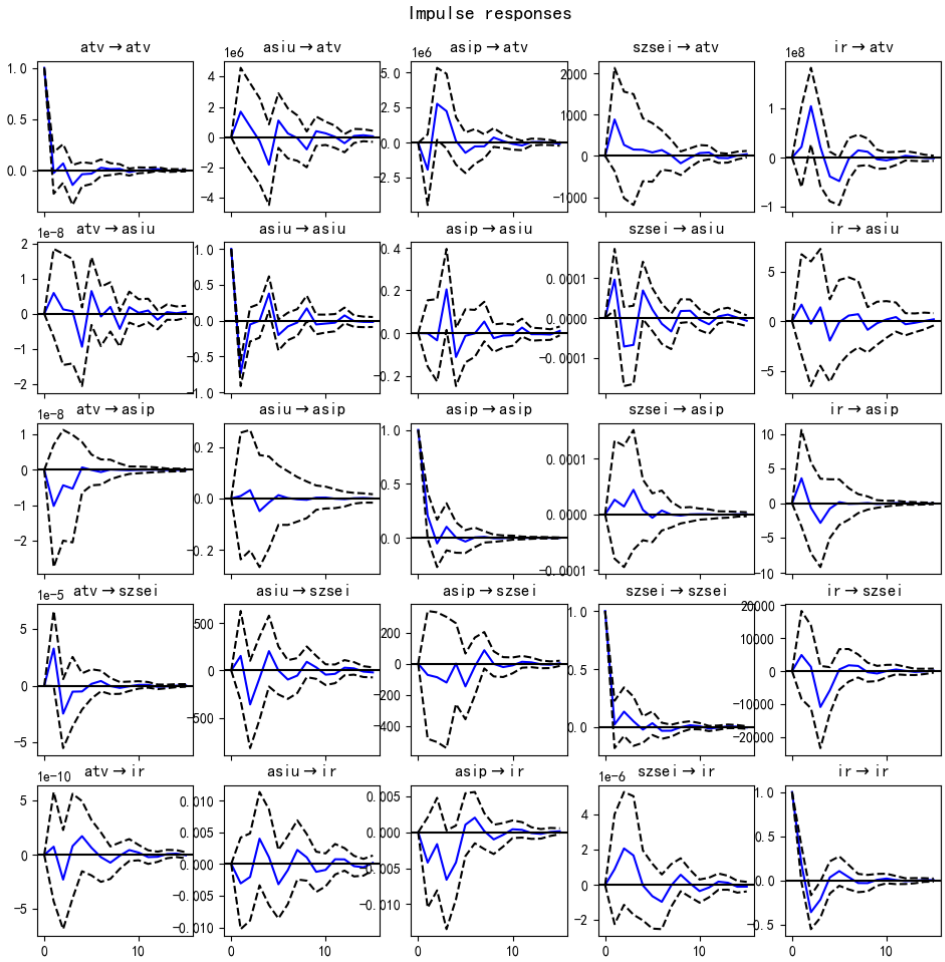

Figure 5.2: Impulse-Responses Figure of Model VAR-ATV

institutional investors' sentiment level can contribute to predicting the change of trading volume.

Based on that, I can also plot the impulse-response analysis figures of model VAR-AI, Model VAR-ASR and model VAR-ATV. The impulse-response function can reflect the dynamic response of the system to the impact from itself or other variables outside. For example, the figure of Model VAR-TV is shown in the Figure 5.2.

The impulse-response figures illustrate different impact patterns of the sen-

timent to the stock price or trading volume. In the model VAR-AI, we can find that both $\Delta ASIU$ and $ASIP$ do not have huge and significant impact on the stock price index change ΔAI. We can tell this from the impulse-response figures of $\Delta ASIU \rightarrow \Delta AI$ and $ASIP \rightarrow \Delta AI$, while the confidence interval is wilder. In the model VAR-ASR, we can find that $\Delta ASIU$ can enhance the stock return in a very short term, but this impact is not so clear. However, in the model VAR-ATV, the impact of $ASIP$ on the change of trading volume ΔATV is obvious and significant. The first instant impact in negative, and then we can observe a reversal positive and maximum impact after term $t+2$. This result aligns the "short-term reversal" phenomenon.

To conclude, the VAR model can provide evidence in the two following conclusions. The first is that the change of individual investors' sentiment can be useful in predicting the stock returns. This implies that when constructing a quantitative trading strategy, the individual user's posts on stock forums are useful indicators. The second is that the total level of institution investors' sentiment can significantly help predicting the change of trading volume. This implies that when building a model to predict the trading volume, institutional investor sentiment can be an important variable.

Generally, align with the previous finding in the OLS model, the information on Guba is "somewhat useful". As we can see, the individual users' posts and the change of their total sentiment level are helpful in predicting the stock returns. However, we cannot find evidence supporting that the institutional users' sentiment is also helpful. This conclusion is quite similar to a previous Chinese research in 2017 (Shi et al., 2017). Their study found evidence that individual and institutional investors' attention can influence the stock return, but

their data showed that the impact of individual investors' attention is greater. Also, same as my finding here, they model also could not find evidence that the sentiment of institutional investors (in their model: news sentiment) can impact the stock price. My model further examines the relationship of sentiment and trading volume. I find strong proof that reversely, the institutional investors' sentiment can influence the trading volume, but I cannot find evidence to support the impact of individual investors' sentiment.

Why do the two sentiment indices have different kinds of impact? Shi et al. provided a convincing explanation of why individual investors' attention can impact the stock return while institutional investors' sentiment cannot. They argued that unlike stock market in developed countries, in Chinese stock market, individual or retail investors are an important part. The extremely high proportion of individual investors make their activity in stock market a mighty force (Shi et al., 2017). This explanation also holds in our data since ChiNext is usually believed to have a higher proportion of individual investors. However, the rise or fall of stock price do not indicate an increase or decrease in trading volume. On the contrary, as we can see from our former descriptive statistics, the sentiment of institutional investors' sentiment is usually neutral. This is determined by the intrinsic of their writing contents, such as newsletter and analysis report. This might be the reason why the institutional investors' sentiment cannot impact the stock return.

5.3 Discussions

5.3.1 Heterogeneous Beliefs Among Individual Investors

Another key concept in behavior finance is the heterogeneous beliefs. Previous empirical studies found that the degree of this heterogeneity will typically rise when the market is transforming from a bullish market to a bearish market. One classical evidence is the rise of heterogeneity in the stock market from 1990s to 2000s when the stock market entered a new round of recession.

Heterogeneous beliefs are nature products and can always be found among individual investors. Economic surveys for investors demonstrated that heterogeneous beliefs can be influenced by the stock market itself (Shefrin and Belotti, 2008). But is the heterogeneity influencing the stock market reversely? The model of Harrison and Kreps predicted yes. They argued that under a "short-sell constraint" (i.e., agent is not allowed to perform short selling), the heterogeneous expectations between agents is going to generate a speculative bubble (Michael Harrison and Kreps, 1978). Later theoretical and empirical research further validated this speculation. Such as the work of Chiarella et al. (Chiarella et al., 2007), which built a more complicated dynamic model in a multi-asset framework; and the work of Boswijk et al. (Boswijk et al., 2007), which used the heterogeneous beliefs to explain the irrational stock price exuberance in the late 1990s.

Given the property of social media, it is easy to estimate the daily heterogeneous beliefs of individual investors. Here, I define the daily heterogeneous beliefs index as the standard error of the sentiment indices of each individual

Table 5.7: Granger Test of the Three Heterogeneity Models

H_0	chi^2	p-Val	Reject
ΔAHB does not Granger-cause ASR	0.91	0.63	FALSE
ASR does not Granger-cause ΔAHB	1.76	0.41	FALSE
ΔAHB does not Granger-cause Delta ATV	0.21	0.90	FALSE
ΔATV does not Granger-cause ΔAHB	0.89	0.64	FALSE
ΔAHB does not Granger-cause Delta $AVOL$	0.61	0.74	FALSE
$\Delta AVOL$ does not Granger-cause ΔAHB	0.10	0.95	FALSE

In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

investors' posts in the following formula.

$$HB_{t,j} = \sqrt{var_i\left(SSI_{i,t}\right)} \tag{5.1}$$

$$AHB_t = \sum_j HB_{t,j} \frac{MV_{t,j}}{\sum_j MV_{t,j}} \tag{5.2}$$

Repeating the VAR model procedures, I can examine whether the daily heterogeneous beliefs index can be useful in predicting the stock return, the trading volume, and the volatility. Here, I report the results of the Granger test results of the three models: VAR-AHB-ASR: daily heterogeneous beliefs index and stock return; VAR-AHB-ATV: daily heterogeneous beliefs index and trading volume; VAR-AHB-AVOL: daily heterogeneous beliefs index and volatility in Table 5.7.

The above Table 5.7 indicates that there is no evidence supporting that ΔAHB can help predicting the stock return, the change of trading volume and the change of volatility. And in turn, the stock return, the change of trading volume and the change of volatility cannot also help predicting the change of daily heterogeneous beliefs index in my dataset.

To conclude, this part I test the Granger causality of ΔAHB with the stock return, the change of trading volume and the change of volatility. I find no evidence supporting that the change of daily heterogeneous beliefs index is useful in predicting stock market. However, this time series data is from a stable stock

market in a short period, i.e., not to bullish or bearish. More similar research should be conducted based on data from a recession or booming stock market.

5.3.2 Firm Size and Sentiment

Firm size is also a key concept in traditional asset pricing theories. Many empirical asset pricing papers use the "size premium" to interpret the role of firm size in asset pricing. And most of the empirical asset pricing works believe the firm size is an important determinant in predicting the expected stock returns (Astakhov et al., 2019). Other global empirical evidence also claimed that firm size has an obvious negative impact on stock returns (Chabachib et al., 2020). But is the sentiment playing a different role in stocks/firms with different sizes?

This discussion part, I examine whether the firm size or firm scale will impact the sensitivity of the stock return, trading volume and volatility facing the impact of investors sentiment. Recall that in Part 3.2, I define a firm/stock is large if its total market value at the end of IPO day is equal to or over 10 billion RMB. I can use this threshold to divide the stocks into groups of large firm size and small firm size. Based on the previous algorithm, I can easily calculate the investors sentiment $LASIU_t$, $LASIP_t$ for large-size stocks and $SASIU_t$, $SASIP_t$ for small size stocks. Following the same notations, I can also define the stock market variables $LASR_t$, $LATV_t$, $LAVOL_t$ for large-size stocks and $SASR_t$, $SATV_t$, $SAVOL_t$ for small-size stocks.

From the above Figure 5.3 of daily sentiments indices of difference size stocks, I can find an obvious difference between stocks of large-size and small-size firms. For large-size firms, individual investors sentiments remain a more

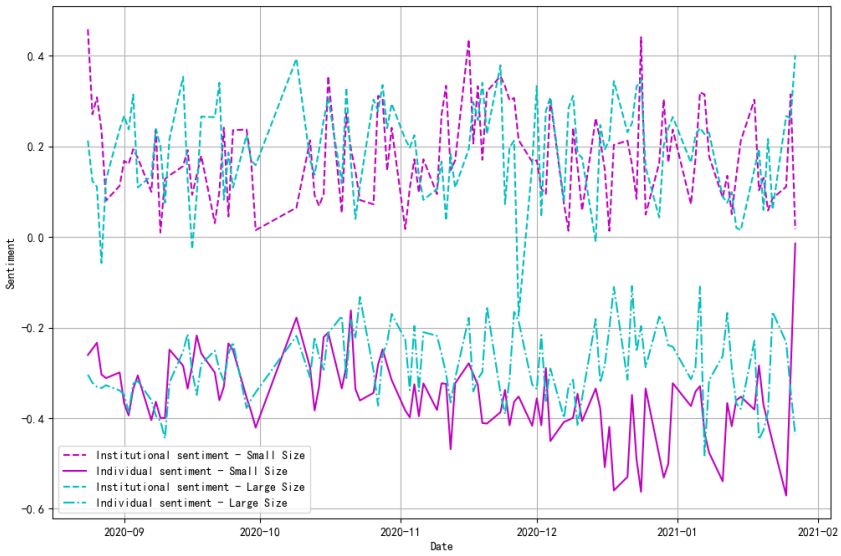

Figure 5.3: Sentiments of Different Size Stocks

stable trend, basically oscillating around -0.2 to -0.3. For smaller size firms, the individual investors sentiment performs a downward trend, and the total sentiment level is far lower than the large-size firms.

This difference might be due to the gap between their stock price indices. The following Figure 5.4 shows a completely distinct trend of the different size firms. We can easily tell that the IPO stock prices for large size firms are more reasonable given their stable tendency around the initial 100 level. For smaller size firms, the IPO prices are set too high, and the stock price is crashing even once lower than 60 level in January 2021.

After the exploratory analysis, I can examine six VAR models here. The VAR-L-ASR, VAR-L-ATV and VAR-L-AVOL are for large-size stocks. The VAR-S-ASR, VAR-S-ATV and VAR-S-AVOL are for small-size stocks. Same as in the Part 5.2, following the standard procedures of VAR models, I can report the

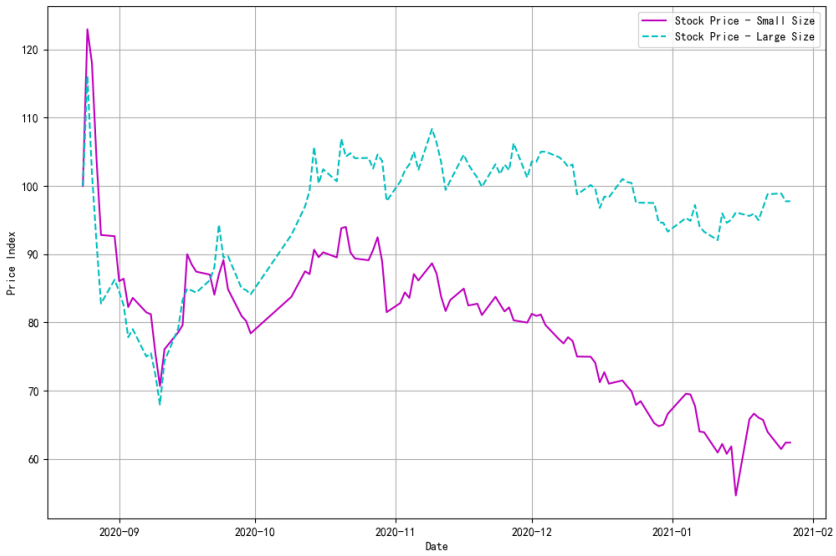

Figure 5.4: Stock Price (Index) of Different Size Stocks

Granger test results in the following Table 5.8.

The results from the Granger test are quite interesting. Focusing on the stocks of large-sized firms, we can observe a similar finding as what we find using integrated model. That is the change of individual investors sentiment can help predicting the stock return. And the institutional investors sentiment can help predicting the trading volume. In large-sized firms, I can also find evidence that the change of individual investors sentiment is useful in volatility models. However, turning our attention to the smaller size firms, I cannot find such empirical evidence. In fact, the VAR models show that the influence of investors sentiment is quite weak on the stock market compared to the influence in larger size firms.

This finding can just my selection of our stock portfolio which combined both large and small firms' stocks. As I reviewed, many previous literature's

Table 5.8: Granger Test Results of Different Firm Sizes

H_0	chi^2	p-Val	Reject
$\Delta LASIU$ does not Granger-cause $LASR$	7.23	0.065*	TRUE
$LASR$ does not Granger-cause $\Delta LASIU$	0.55	0.91	FALSE
$LASIP$ does not Granger-cause $LASR$	1.79	0.62	FALSE
$\Delta LASIU$ does not Granger-cause $\Delta LATV$	2.69	0.44	FALSE
$LASIP$ does not Granger-cause $\Delta LATV$	7.53	0.06*	TRUE
$\Delta LATV$ does not Granger-cause $LASIP$	4.08	0.25	FALSE
$\Delta LASIU$ does not Granger-cause $\Delta LAVOL$	7.24	0.065*	TRUE
$\Delta LAVOL$ does not Granger-cause $\Delta LASIU$	3.27	0.35	FALSE
$LASIP$ does not Granger-cause $\Delta LAVOL$	2.67	0.45	FALSE
$\Delta SASIU$ does not Granger-cause $SASR$	1.25	0.74	FALSE
$SASIP$ does not Granger-cause $SASR$	2.27	0.52	FALSE
$\Delta SASIU$ does not Granger-cause $\Delta SATV$	3.98	0.26	FALSE
$SASIP$ does not Granger-cause $\Delta SATV$	0.67	0.88	FALSE
$\Delta SASIU$ does not Granger-cause $\Delta SAVOL$	2.88	0.41	FALSE
$SASIP$ does not Granger-cause $\Delta SAVOL$	0.10	0.99	FALSE

In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

work and finding are based on index biased to large companies listed, such as S&P 500 in US stock market and CSI 300 in Chinese stock market. If the firm size can indeed influence the impact of sentiment, their conclusion on investors sentiment could be exaggerated. However, given the length limitation of my time series data, more cross-sectional research projects are encouraged to study the impact of sentiment using data from stocks of different firm sizes.

To conclude, this part I examine the possible relationship of sentiment impact and firm size. My empirical study shows that the market and the sentiment of large and small size firms have clearly different trends. My VAR models shows that the sentiment is useful in predicting the large firms' stocks. Nevertheless, no such evidence is found to prove that the sentiment is also helpful in predicting the small firms' stocks.

CHAPTER 6

CONCLUSIONS

Focusing on the new IPO registration system of Chinese stock market, I choose Guba as my target social media platform and uses big data techniques to gather the posts data of stocks whose IPOs are based on this new registration system. Then, I use NLP algorithms and a pre-trained sentiment analysis model to construct the sentiment proxy index for investor sentiments from the posts data. Next, I split the sentiment index into individual investors' sentiment index and institutional investors' sentiment index. After that,I study their influence on stock market respectively using simple OLS regression and VAR models.

In this thesis, I find that if considering the stock portfolio as an entirely, the investors' sentiment can indeed help predicting the stock market. But different kinds of investors' sentiment have completely different kinds of influence. For individual investors, my VAR models provide evidence that their sentiment can help predicting the stock return. But I have not found such evidence that their sentiment can be useful in predicting the trading volume. For institutional investors, my VAR models indicates that their sentiment can be a significantly important determinant in models of trading volume. But the VAR models do not support the claim that their sentiment can also be such a determinant in models of stock return.

In addition, in this thesis I expand my discussion to two salient topics in traditional and behavior asset pricing. The first topic is the heterogeneous beliefs. This thesis uses the standard errors of daily posts sentiment as the proxy of heterogeneous beliefs. My VAR models shows that I cannot find proofs that my proxy of heterogeneous beliefs can impact the stock market. However, this

conclusion is found from times series data from a stable stock market in a short period. More research projects are encouraged to investigate the impact of heterogeneous beliefs during recession or blooming periods.

The second topic is the firm size. Here, I use 10 billion RMB market value as the threshold to divide the firms into large size firms and small size firms. After following the same procedure, I study the impact of sentiment for both large size firms and small size firms. In large size firms' group, the VAR models provide similar evidence: the change of individual investors sentiment can help predicting the stock return as well as the change of stock volatility, and the institutional investors sentiment can help predicting the change of trading volume. In small size firms' group, the VAR models cannot provide the same evidence. In fact, my results shows that the impact of investors sentiment in small size firms' group is limited.

This thesis provides two new directions for further studies. The first one is the firm size problem. Many previous similar research projects are based on index obviously biased to large, listed companies, such as S&P 500 in US stock market and CSI 300 in Chinese stock market. The empirical finding of this thesis challenges this kind of stock portfolio, since it might exaggerate the impact of investors' sentiment. Later cross-sectional research can study the impact of sentiment using longer data from stocks of different firm sizes. Another direction of further research can be the impact of regulation. For instance, can the limit to arbitrage ease or magnify the impact of investors' sentiment? Or can the reformed registration IPO system in Chinese stock market ease or magnify the impact of investors' sentiment? Cross-sectional studies on regulation and sentiment are welcomed to amplify the research on investors' sentiment.

APPENDIX A

APPENDIX: MISCELLANEOUS TABLES

A.1 Descriptive Statistics of Data Before Average

The following tables Table A.1 and A.2 are the descriptive statistics of data before average, including the sentiment indices and stock market data.

Note that in these tables, the sentiment indices are not filtered yet. This means the sentiment indices are not all from trading days. Instead, some of the indices are generated from posts on weekends or holidays.

A.2 OLS Results Tables of Model 4.3 and 4.6

The following tables Table A.3 and A.4 are the OLS regression results of Model 4.3 and 4.6.

Table A.1: Descriptive Statistics of Sentiment Indices

Variable	Min	Max	Mean	Std. Dev.	Kurtosis	Skewness
SIP_t	-0.971	0.999	0.162	0.303	1.480	0.720
SIU_t	-0.993	0.999	-0.278	0.274	2.198	0.855

Table A.2: Descriptive Statistics for Stock Market Data

	Unit	Min	Max	Mean	Std. Dev	Kurtosis	Skewness
CP_t	RMB	28	208	88.48	44.17	-0.63	0.70
TV_t	/	275083	44999376	5549773.59	5851902.45	3.63	1.79
TA_t	RMB	25349348.58	3265937281	462806391.70	560471526.70	6.17	2.32
TR_t	%	1.32	76.09	19.46	18.78	0.40	1.30
SR_t	/	-0.25	0.45	-0.002	0.04	17.66	0.93
VOL_t	/	0	0.12	0.03	0.02	0.92	0.63

Table A.3: Result of Model 4.3

	(1) $ASR(t)$	(2) $ASR(t+1)$	(3) $ASR(t+2)$
Constant	0.617***	0.525***	0.502***
	(0.13)	(0.13)	(0.12)
$ASIU$	0.110**	0.0360	-0.0858**
	(0.04)	(0.043)	(0.042)
$ASIP$	0.0213	-0.0227	-0.0350
	(0.03)	(0.032)	(0.031)
$AVOL$	-1.41***	-1.335***	-0.8324*
	(0.51)	(0.503)	(0.494)
$SZSEI$	-1.803e-5***	-2.022e-05***	-2.117e-05***
	(5.96e-06)	(5.86e-06)	(5.76e-06)
ΔEPU	0.0001***	7.037e-05**	9.708e-05***
	(3.58e-05)	(3.51e-05)	(3.44e-05)
IR	0.7944	0.8959	-0.3635
	(0.610)	(0.617)	(0.623)
$AI(t-1)$	-0.0024***	-0.0007	-0.0032***
	(0.001)	(0.001)	(0.001)
$AI(t-2)$	2.045e-05	-0.0036***	0.0025**
	(0.001)	(0.001)	(0.001)
$AI(t-3)$	-0.0009	0.0021***	-0.0014*
	(0.001)	(0.001)	(0.001)
Adj. R^2	0.257	0.212	0.161

This table presents OLS regression result of relation of individual and institutional investor's sentiment with stock return. I also relate the stock price in the future two days to $ASIU$ and $ASIP$. The control variables are the lagged values of AI up to three days, the volatility AVOL, the total index of Shenzhen Exchange $SZSEI$, the change of economic policy uncertainty index ΔEPU and the interest rate IR. In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

Table A.4: Result of Model 4.6

	(1) $ATR(t)$	(2) $ATR(t+1)$	(3) $ATR(t+2)$
Constant	-36.5744*	22.4598	60.2722***
	(18.085)	(19.179)	(18.552)
$ASIU$	4.9008	8.6219	6.7045
	(6.138)	(6.522)	(6.286)
$ASIP$	2.2055	-3.2475	3.0590
	(4.507)	(4.841)	(4.686)
$AVOL$	835.4674***	623.1234	462.7300***
	(72.297)	(77.001)	(74.771)
$SZSEI$	0.0009	-0.0012	-0.0027***
	(0.001)	(0.001)	(0.001)
ΔEPU	0.0019	0.0078	0.0103*
	(0.005)	(0.005)	(0.005)
IR	-218.4893**	-116.7946	-29.3735
	(86.376)	(94.493)	(94.169)
$AI(t-1)$	0.4040**	0.1516	-0.1043
	(0.156)	(0.165)	(0.159)
$AI(t-2)$	0.1569	0.1129	0.1044
	(0.165)	(0.175)	(0.169)
$AI(t-3)$	-0.3892***	-0.3412***	-0.2479*
	(0.113)	(0.120)	(0.116)
Adj. R^2	0.716	0.644	0.638

This table presents OLS regression result of relation of individual and institutional investor's sentiment with turnover rate. I also relate the stock price in the future two days to $ASIU$ and $ASIP$. The control variables are the lagged values of AI up to three days, the volatility AVOL, the total index of Shenzhen Exchange $SZSEI$, the change of economic policy uncertainty index ΔEPU and the interest rate IR. In this table, ***, ** and * represent 1%, 5% and 10% level of significance, respectively.

APPENDIX B

APPENDIX: MISCELLANEOUS FIGURES

B.1 Introduction to SKEP

Here is a brief introduction to the SKEP algorithm. The Figure B.1 is from the work of Tian et al. from Baidu Inc., & USTC (Tian et al., 2020).

The first part is sentiment masking. It recognizes the sentiment information of an input sequence based on automatically-mined sentiment knowledge and produces a corrupted version by removing these information.

The second part is known as the sentiment pre-training. The objectives require the transformer to recover the removed information from the corrupted version.

B.2 ACF Plots for Residuals

The following plots Figure B.2 and B.3 are the ACF Plots for residuals of model VAR-ASR and VAR-ATV.

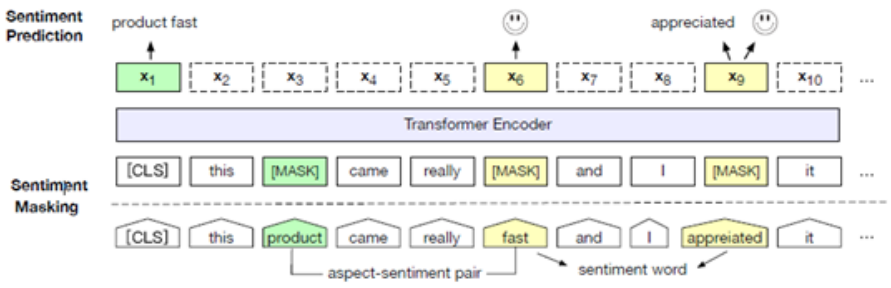

Figure B.1: Figure of the Working Mechanism of SKEP

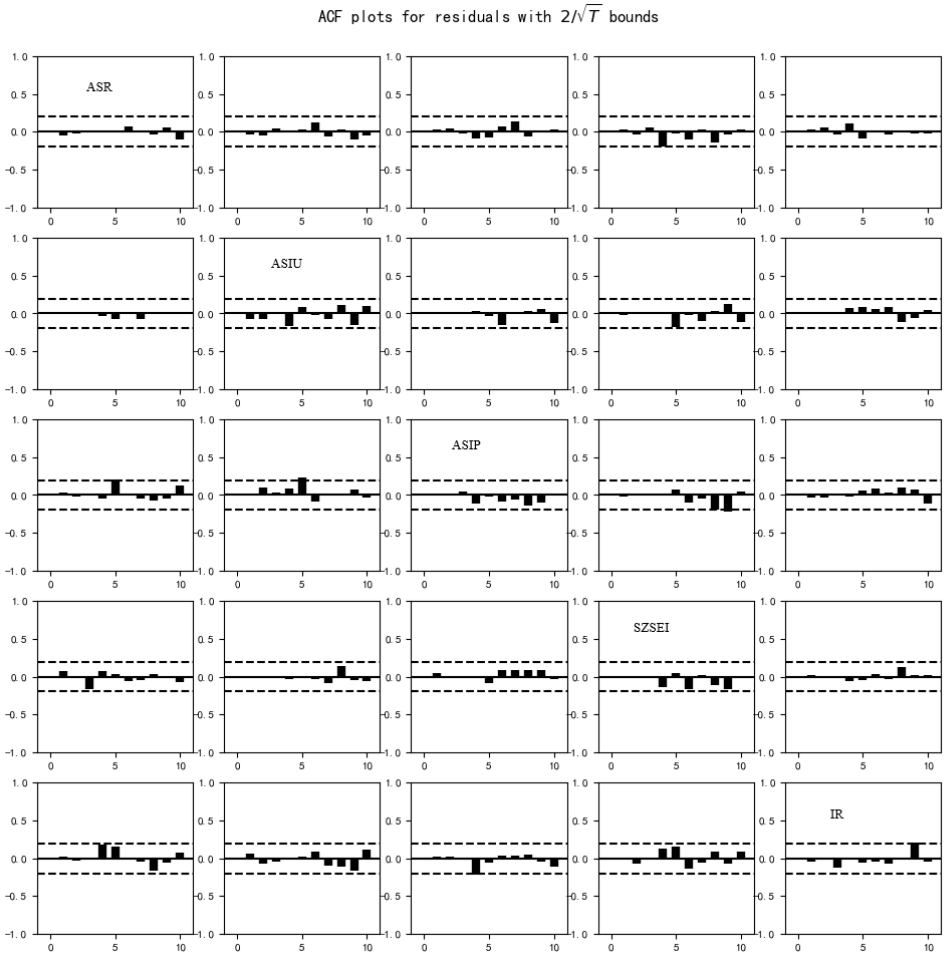

Figure B.2: ACF Plots for Residuals - Model VAR-ASR

B.3 Impulse-Responses Figures

The following plots Figure B.4 and B.5 are the impulse-response plots for model VAR-AI and VAR-ASR.

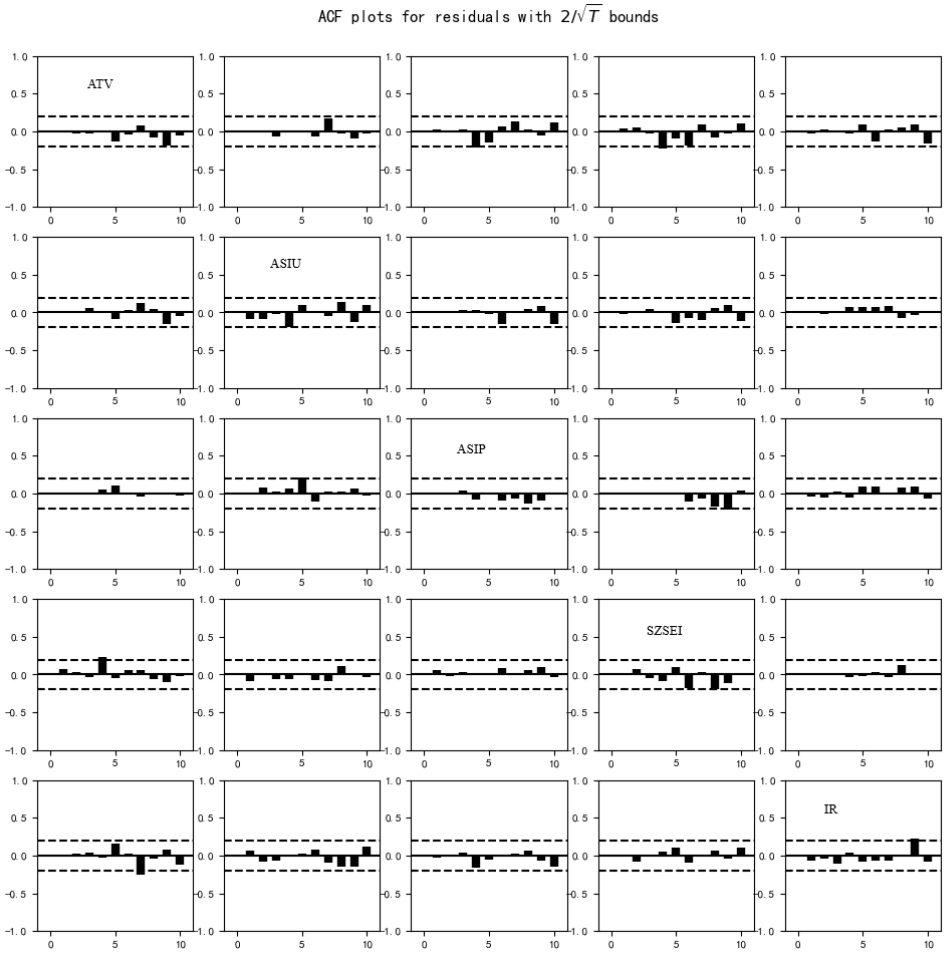

Figure B.3: ACF Plots for Residuals - Model VAR-ATV

B.4 Daily Heterogeneous Beliefs Index and Stock Return

The following plot Figure B.6 decribes the possible relationship of the heterogeneous beliefs and the avergae stock return.

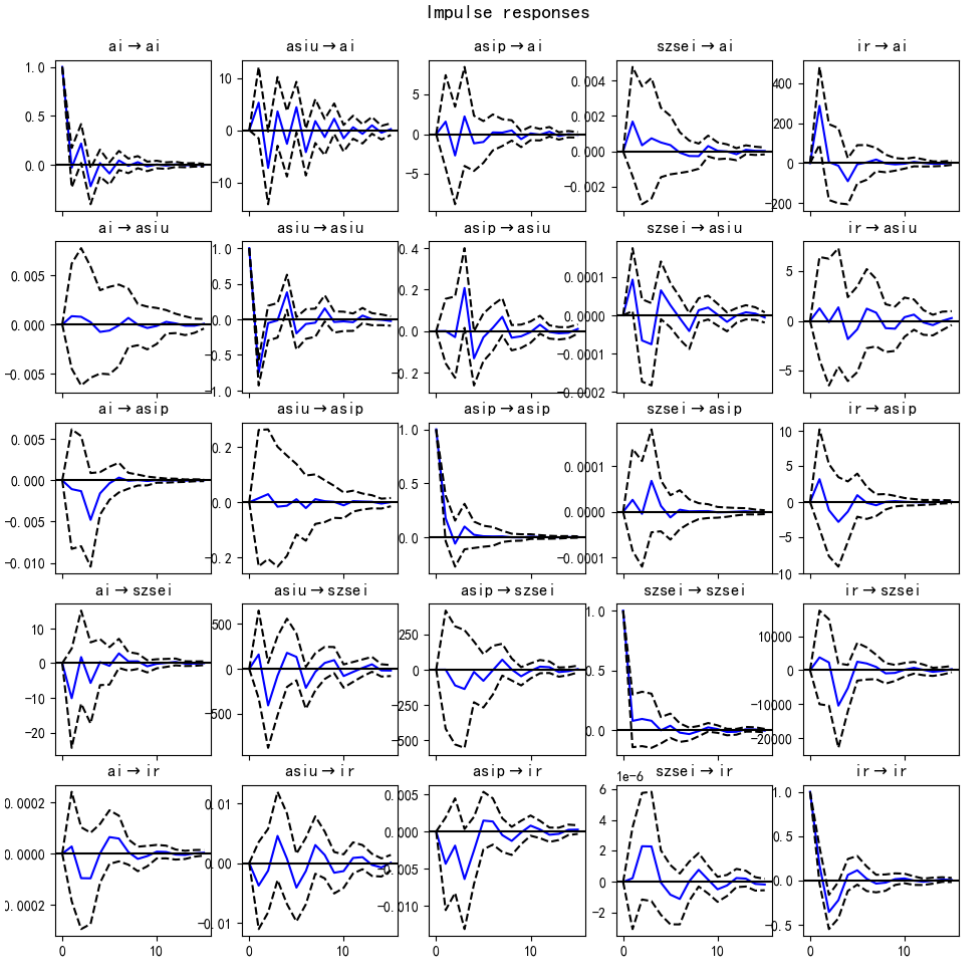

Figure B.4: Impulse-Responses Figure of Model VAR-AI

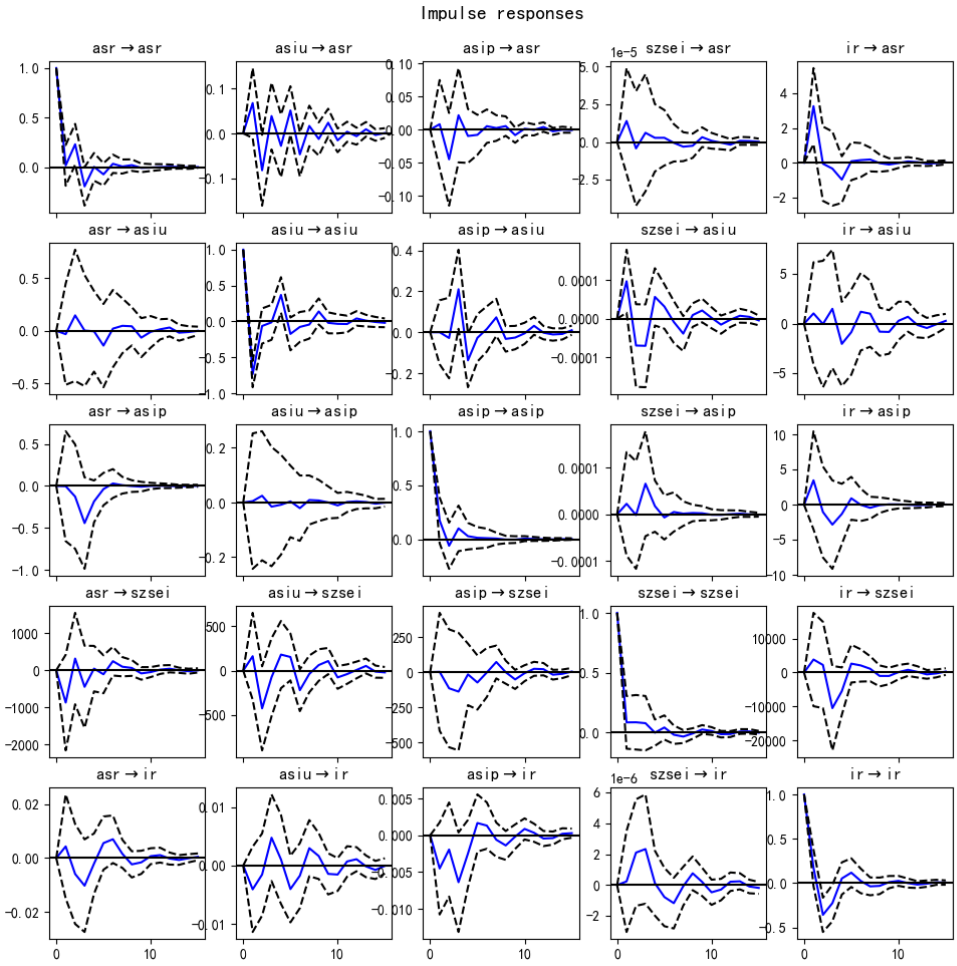

Figure B.5: Impulse-Responses Figure of Model VAR-ASR

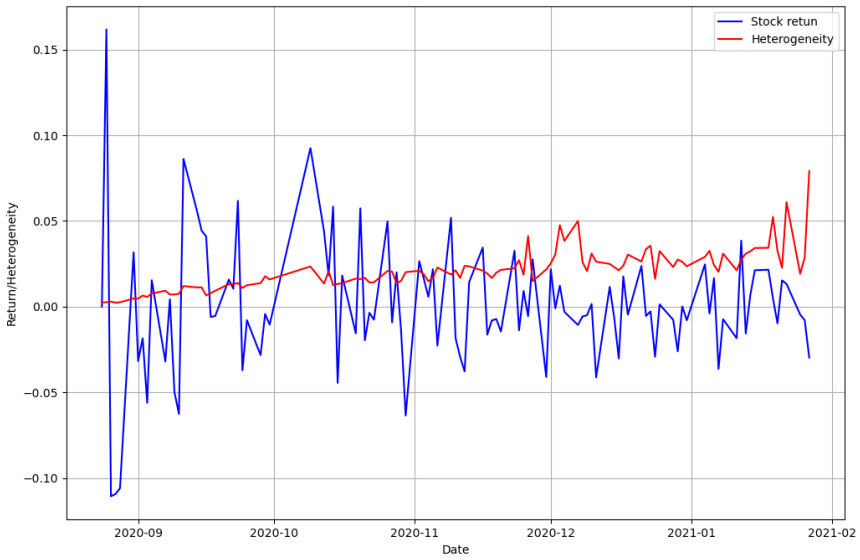

Figure B.6: Daily Heterogeneous Beliefs Index and Stock Return

BIBLIOGRAPHY

Jawad M. Addoum and Alok Kumar. Political sentiment and predictable returns. *Review of Financial Studies*, 29(12):3471–3518, 2016. ISSN 14657368. doi: 10.1093/rfs/hhw066.

Omar Y. Adwan, Marwan Al-Tawil, Ammar M. Huneiti, Rawan A. Shahin, Abeer A. Abu Zayed, and Razan H. Al-Dibsi. Twitter sentiment analysis approaches: A survey. *International Journal of Emerging Technologies in Learning*, 15(15):79–93, 2020. ISSN 18630383. doi: 10.3991/ijet.v15i15.14467.

Manuel Ammann and Nic Schaub. Do Individual Investors Trade on Investment-Related Internet Postings? *Management Science*, 2020. ISSN 0025-1909. doi: 10.1287/mnsc.2020.3733.

Anton Astakhov, Tomas Havranek, and Jiri Novak. FIRM SIZE AND STOCK RETURNS: A QUANTITATIVE SURVEY. *Journal of Economic Surveys*, 33(5), 2019. ISSN 14676419. doi: 10.1111/joes.12335.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 2006. ISSN 00221082. doi: 10.1111/j.1540-6261.2006.00885.x.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. In *Journal of Economic Perspectives*, volume 21, 2007. doi: 10.1257/jep.21.2.129.

Malcolm Baker, Jeffrey Wurgler, and Yu Yuan. Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2), 2012. ISSN 0304405X. doi: 10.1016/j.jfineco.2011.11.002.

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343, sep 1998. ISSN 0304405X. doi: 10.1016/s0304-405x(98)00027-0.

Shibo Bian, Dekui Jia, Feng Li, and Zhipeng Yan. A New Chinese Financial Sentiment Dictionary for Textual Analysis in Accounting and Finance. *SSRN Electronic Journal*, (February 2018):1–21, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3446388.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language ToolKit (NLTK) Book*. 2009.

H. Peter Boswijk, Cars H. Hommes, and Sebastiano Manzan. Behavioral heterogeneity in stock prices. *Journal of Economic Dynamics and Control*, 31(6), 2007. ISSN 01651889. doi: 10.1016/j.jedc.2007.01.001.

Gregory W. Brown and Michael T. Cliff. Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1), 2004. ISSN 09275398. doi: 10.1016/j.jempfin.2002.12.001.

Gregory W. Brown and Michael T. Cliff. Investor sentiment and asset valuation. *Journal of Business*, 78(2), 2005. ISSN 00219398. doi: 10.1086/427633.

Jaroslav Bukovina. Social media big data and capital markets-An overview, sep 2016. ISSN 22146369.

Erik Cambria and Bebo White. Jumping NLP curves: A review of natural language processing research, 2014. ISSN 1556603X.

Mochammad Chabachib, Ike Setyaningrum, Hersugondo Hersugondo, Intan Shaferi, and Imang Dapit Pamungkas. Does financial performance matter?

Evidence on the impact of liquidity and firm size on stock return in Indonesia. *International Journal of Financial Research*, 11(4), 2020. ISSN 19234031. doi: 10.5430/ijfr.v11n4p546.

Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 1, pages 419–429. Association for Computational Linguistics (ACL), 2015. ISBN 9781941643723. doi: 10.3115/v1/p15-1041.

Carl Chiarella, Roberto Dieci, and Xue Zhong He. Heterogeneous expectations and speculative behavior in a dynamic multi-asset framework. *Journal of Economic Behavior and Organization*, 62(3):408–427, 2007. ISSN 01672681. doi: 10.1016/j.jebo.2005.08.005.

Zhi Da, Joseph Engelberg, and Pengjie Gao. The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, 28(1), 2015. ISSN 14657368. doi: 10.1093/rfs/hhu072.

Sanjiv R. Das and Mike Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the Web. *Management Science*, 53(9):1375–1388, sep 2007. ISSN 00251909. doi: 10.1287/mnsc.1070.0704.

J. Bradford De Long, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. Noise Trader Risk in Financial Markets. *Journal of Political Economy*, 98(4), 1990. ISSN 0022-3808. doi: 10.1086/261703.

Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based ap-

proach to opinion mining. *WSDM08 - Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–239, 2008. doi: 10.1145/1341531.1341561.

Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 2008. ISSN 15487660. doi: 10.18637/jss.v025.i05.

Kenneth L. Fisher and Meir Statman. Investor Sentiment and Stock Returns. *Financial Analysts Journal*, 56(2):16–23, 2000. ISSN 0015198X. doi: 10.2469/faj.v56.n2.2340.

John J. Hartman, Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvia. The General Inquirer: A Computer Approach to Content Analysis. *American Sociological Review*, 32(5), 1967. ISSN 00031224. doi: 10.2307/2092070.

Dashan Huang, Fuwei Jiang, Jun Tu, and Guofu Zhou. Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies*, 28(3), 2015. ISSN 14657368. doi: 10.1093/rfs/hhu080.

Yumei Jiang and Mingzhao Wang. Investor Sentiment and Stock Returns: An Empirical Study on Aggregate Effects and Cross-section Effects. *Nankai Business Review*, (03):150–160, 2010.

Kissan Joseph, M. Babajide Wintoki, and Zelin Zhang. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, 27(4), 2011. ISSN 01692070. doi: 10.1016/j.ijforecast.2010.11.001.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 1, 2014. doi: 10.3115/v1/p14-1062.

Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 2011. ISSN 00076813. doi: 10.1016/j.bushor.2011.01.005.

Alok Kumar and Charles M.C. Lee. Retail investor sentiment and return comovements. *Journal of Finance*, 61(5), 2006. ISSN 00221082. doi: 10.1111/j.1540-6261.2006.01063.x.

Jia Li, Yun Chen, Yan Shen, Zhuo Huang, and Jingyi Wang. Measuring China's Stock Market Sentiment. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3377684.

Lei Li. The Influence of Government Regulation on IPO Underpricing. *Technology and Investment*, 09(02):109–116, 2018. ISSN 2150-4059. doi: 10.4236/ti.2018.92008.

Zhongguo Li and Maosong Sun. Punctuation as implicit annotations for Chinese word segmentation. *Computational Linguistics*, 35(4), 2009. ISSN 15309312. doi: 10.1162/coli.2009.35.4.35403.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. ISSN 23318422.

M. Ángeles López-Cabarcos, Ada M. Pérez-Pico, Paula Vázquez-Rodríguez, and M. Luisa López-Pérez. Investor sentiment in the theoretical field of behavioural finance. *Economic Research-Ekonomska Istrazivanja*, 33(1), 2020. ISSN 1331677X. doi: 10.1080/1331677X.2018.1559748.

Tim Loughran and Bill McDonald. IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, 109(2), 2013. ISSN 0304405X. doi: 10.1016/j.jfineco.2013.02.017.

Mika V Mantyla, Daniel Graziotin, and Miikka Kuutila. The Evolution of Sentiment Analysis - A Review of Research Topics , Venues , and Top Cited Papers. 27(February):16–32, 2018.

Zachary McGurk, Adam Nowak, and Joshua C. Hall. Stock returns and investor sentiment: textual analysis and social media. *Journal of Economics and Finance*, 44(3), 2020. ISSN 19389744. doi: 10.1007/s12197-019-09494-4.

J. Michael Harrison and David M. Kreps. Speculative Investor Behavior In A Stock Market With Heterogeneous Expectations. *Quarterly Journal of Economics*, 92(2), 1978. ISSN 15314650. doi: 10.2307/1884166.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Technical report, 2002. URL `http://www.cs.cornell.edu/people/pabo/movie-review-data/`.

Haiyun Peng, Erik Cambria, and Amir Hussain. A Review of Sentiment Analysis Research in Chinese Language, aug 2017. ISSN 18669964.

Thomas Renault. Sentiment analysis and machine learning in finance: a com-

parison of methods and models on one million messages. *Digital Finance*, 2 (1-2), 2020. ISSN 2524-6984. doi: 10.1007/s42521-019-00014-x.

Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management*, 52(1):5–19, jan 2016. ISSN 03064573. doi: 10.1016/j.ipm.2015.01.005.

Maik Schmeling. Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, 16(3), 2009. ISSN 09275398. doi: 10.1016/j.jempfin.2009.01.002.

Hersh Shefrin and Mario L. Belotti. *A Behavioral Approach to Asset Pricing*. 2008. ISBN 9780123743565. doi: 10.1016/B978-0-12-374356-5.X5001-3.

Yong Shi, Jing Tang, and Kun Guo. The Study of Social Media Investor Attention and Sentiment's Influence on Chinese Stock Market. *Journal of Central University of Finance and Economics*, (07):45–53, 2017.

Robert J. Shiller. From efficient markets theory to behavioral finance. In *Journal of Economic Perspectives*, volume 17, 2003. doi: 10.1257/089533003321164967.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 2011. ISSN 15309312. doi: 10.1162/COLI-a-00049.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv*, 2020. ISSN 23318422. doi: 10.18653/v1/2020.acl-main.374.

Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. dec 2002. URL `https://arxiv.org/abs/cs/0212032`.

Xiaojun Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL*, pages 553–561. Association for Computational Linguistics (ACL), 2008. doi: 10.3115/1613715.1613783.

Hongbo Yi, Juanjuan Lai, and Dayong Dong. A Study of the Influence of BBS Investor Sentiments on the Trading Market——An Empirical Analysis Based on VAR Model. *Collected Essays on Finance and Economics*, 01(01):46–54, 2015. doi: 10.13762/j.cnki.cjlc.2015.01.007.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. HHMM-based Chinese lexical analyzer ICTCLAS. 2003. doi: 10.3115/1119250.1119280.

Qiang Zhang and Shu'e Yang. Noise trading,investor sentiment volatility and stock returns. *Systems Engineering-Theory and Practice*, 03(03):40–47, 2009.

Qiang Zhang, Shu'e Yang, and Hong Yang. An Empirical Study on Investors' Sentiment and Stock Returns in Chinese Stock Market. *Systems Engineering*, 07(07):13–17, 2007.

Yihao Zhang, Yuan Li, Zhongfeng Su, and Zelin Zhang. Can Internet Search Predict the Stock Market? *Journal of Financial Research*, 02(02):193–206, 2014.