LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

# MASTER

MASTER IN FINANCE

# MASTER´S FINAL WORK

DISSERTATION

# ANALYSIS OF THE RELATIONSHIP BETWEEN THE SENTIMENT OF RETAIL INVESTORS AND THE PERFORMANCE OF THE CHINESE STOCK MARKET

YONGZHE ZHAO

JULY- 2020

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

# MASTER
MASTER IN FINANCE

# MASTER´S FINAL WORK
DISSERTATION

# ANALYSIS OF THE RELATIONSHIP BETWEEN THE SENTIMENT OF RETAIL INVESTORS AND THE PERFORMANCE OF THE CHINESE STOCK MARKET

YONGZHE ZHAO

**SUPERVISION:**
PROFESSOR DOUTOR PEDRO VERGA MATOS

JULY-2020

## GLOSSARY

TLCC – Time lag cross correlation.

VAR – Vector autoregression model.

ABSTRACT, KEYWORDS AND JEL CODES

Unlike stock markets in developed countries, Chinese stock markets are mainly composed of retail investors. Retail investment behavior is susceptible to emotions, which can affect the performance of stock markets. By studying the relationship between the two types of stock markets, retail investors can increase their awareness of risk and rational investment, and the regulation of Chinese capital markets can also be developed more scientifically and healthily. In this paper, the affective computing method is used to quantify the sentiment of retail investors registered on the Shanghai Stock Exchange. Then, the retail sentiment time series, the closing price of the Shanghai Securities Composite Index, and the total trading volume of the Shanghai Stock Exchange are organized for analysis and assessed through three analysis methods, the VAR model, Pearson correlation, and TLCC. The conclusions drawn from this study are as follows: (i) There is no causal relationship between the sentiment of retail investors and the closing price of the Shanghai Securities Composite Index. (ii) There is a causal relationship between retail investor sentiment and the total trading volume of the Shanghai Stock Exchange. (iii) There is a mutual lag influence and strong correlation between the sentiment of retail investors and the changing rate of the Shanghai Securities Composite Index.

KEYWORDS: Behavioral Finance; Crawler; Affective Computing; Sentiment Dictionary Method; Pearson Correlation; TLCC; VAR; Shanghai Securities Composite Index.

JEL CODES: C23; C36; D24; D43; E32; L22.

TABLE OF CONTENTS

TABLE OF FIGURES

## ACKNOWLEDGMENTS

There are times when I would close my eyes and start to recall the difficult journey of the last two years. Everything goes through my mind. I feel very emotional, and I calm down again. It is time to say “goodbye”. I regard studying at ISEG as a turning point in my life, and it truly has been.

I would like to express my sincere thanks to my supervisor Pedro Verga Matos for his guidance during this special pandemic period. Without his help, I would still be wandering in confusion.

I would like to thank ISEG for providing a good study environment. Thank you ISEG for each time you defended my rights during my exchange student time.

I would like to thank Portugal and the EU. I truly appreciate that I can experience the diverse culture and values. It is from here that I found what I truly appreciate and where I want to go.

Finally, I am also grateful to all my professors, classmates, and friends. They composed every moment in my life at ISEG, and they also composed each fragment of happiness here.

# ANALYSIS OF THE RELATIONSHIP BETWEEN THE SENTIMENT OF RETAIL INVESTORS AND THE PERFORMANCE OF THE CHINESE STOCK MARKET

By Yongzhe Zhao

## 1. INTRODUCTION

### *1.1. Research background of the subject*

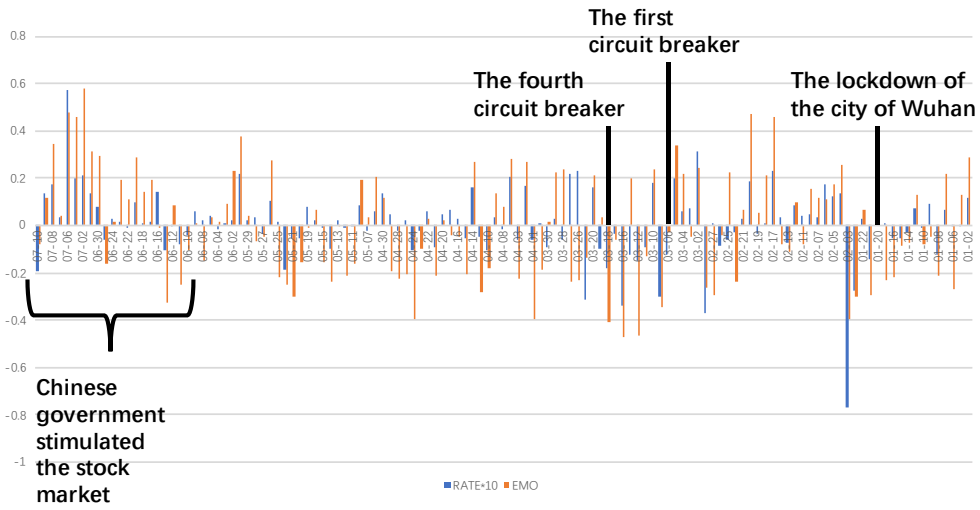
With the development of technology, especially IT technology, the method of acquiring the investment information and connections of retail investors has been reformed in recent years. Retail investors share their opinions on public internet media, which creates a large amount of unstructured data. Furthermore, retail investors are also influenced by the opinions and subjective sentiment shared by other retail investors. Mei et al. (2009) and Carpenter et al. (2015) document that the Chinese stock markets are highly speculative and dominated by inexperienced retail investors who are subject to investor psychological biases. However, in the US, institutional investors dominate the market. Therefore, the relationship between stock market performance and the sentiment movement of Chinese retail investors can be observed in the Chinese stock markets.

In the past, because of the limited technology available to analyze unstructured data, it was difficult to scientifically perform effective computing to measure the sentiment of retail investors. As substantial progress has been achieved in data science, collecting, organizing and analyzing unstructured data are more effective and quicker than they used to be. Data mining methods enable the value of unstructured data to be discovered. Combining sentiment unstructured data analysis with the background of the majority of retail investors in the Chinese stock markets may provide valuable analysis results.

From 01/01/2020 to 10/07/2020, several events destroyed the confidence of retail investors in the Chinese capital markets. Additionally, some events increased the confidence of retail investors and the performance of markets at the same time. Zhang et al (2020) showed that the lockdown of the entire city of Wuhan on 23 January 2020 shocked the whole world and later proved to be a very effective policy intervention by the Chinese government. For example, the investigation of Huo and Qiu (2020) shows that overreactions are stronger for stocks with lower institutional ownership, which means that retail investors reacted more strongly to COVID-19. The conflict between the two

major economies of China and the United States has intensified. The research of Wang et al (2020) shows that the negative effect on stocks is stronger for firms with prior export exposure to the US, especially nonstate firms. On March 22, the circuit breaker of the U.S. stock market was triggered four times. Xuan (2020) shows that the US stock market had significant spillover effects on the returns and fluctuations of the Chinese stock markets in the circuit breaker periods.

FIGURE 1 – Affective Computing Emotion Results of retail investors (EMO) and the Shanghai Securities Composite Index Growth Rate (RATE).




In June, the Chinese government began to consciously guide public opinion to promote stock market reform and long-term development in a healthier way by launching a package of open systems and reforms of the Chinese stock market and spreading hints by previous high officials to the market. Public sentiment was ignited, which immediately brought prosperity to the stock market. These abovementioned events are reflected in Figure 1. Figure 1 also shows that with the extreme dynamic change of sentiment, the performance of the capital market also experienced extremely dynamic changes compared to what was normal. Fang et al. (2019) showed that through the "alarm effect1" and "herd effect2", emerging markets were infected by the residual contagion of the international financial crisis due to their weak macro foundations and immature financial

[1] The alarm effect means that when dangerous signals emerge in a market, investors sell in an irrational way.

[2] The herd effect means that investors engage in blind obedience behavior in the market.

systems. Regarding these two intense negative events, the “herd effect” in this last half year and the relationship between the emotions of retail investors and the performance of the capital markets should be studied.

### *1.2.Research method*

In the past, structured data were used for affective computing and quantizing the sentiment of retail investors. These data include macroeconomic indicators (such as the consumer confidence index) and the transaction records of capital markets. For example, Ben-Rephae et al (2012) use mutual fund flows to measure sentiment. Misina (2003) uses the risk appetite index to measure sentiment. Whether these structured data can directly reflect the relative sentiment level is still a concern. However, unstructured data are not mainly used for analysis. Unstructured data, for example, comments on social media, such as Twitter and discussions about news and topics on public news websites and retail investors’ internet forums, are used by retail investors to express their feelings and thoughts. The difficulty is how to design a system that can acquire data automatically from the target web or platform and then clean, code, and quantize sentiment for further analysis. The most famous internet forums of the Chinese Shanghai Securities Composite Index are a source of data. Many retail investors gather together there for discussions and sharing their opinions. A crawler algorithm programmed in Python language was used to automatically obtain, clean, and reorganize these unstructured data. Then, the Jieba[3] method was used to cut sentences based on the grammar and tradition of the Chinese language and into words. Then, the affective computing algorithm based on the sentiment dictionary was used to quantize sentiment and export the results. The relationship between individual events and sentiment results was analyzed. Next, EViews was used to build the VAR model for the Granger causal relation test and impulse response analysis. Then, the Pearson correlation and time lag cross-correlation analyses were used to analyze the relationship of the time series. Furthermore, other methods were used in the analysis.

[3] The Jieba method is a famous and widely used Chinese text segmentation library based on Python.

## 2.Literature research

### *2.1. Sentiment and Investors' behaviour*

The literature research was mainly composed of the following parts: (i) Emotions and investors' behavior, (ii) behavioral finance and efficient market theory, and (iii) Chinese stock markets.

In his paper, Peterson (2007) mentions that several studies have proven that individual investors' financial decisions are significantly influenced by their emotions and moods. This is the foundation of affective computing in stock markets. Sun et al (2016) found that when the economy is booming and the market volume is large, emotions are much more predictable for markets. However, during a recession, emotions are less predictable for markets, and during periods of low trading volume, the predictability is greatly reduced.

The research of Tsai (2017) revealed that the contagion of investor sentiment is asymmetric. When institutional investors remain optimistic and the market performs well, the spread of sentiment among investors is not obvious. Conversely, pessimism spreads more easily.

In the past, the measurement of investors' sentiment relied on traditional surveys or structured data. Comparing different traditional measures, it is common to find that the sentiment results are quite different from each other. Qiu and Welch (2004) stated that the traditional method may have difficulty measuring sentiment. Because of equipment and data mining restrictions, Wüthrich et al (1998) use the simple terms frequency[4], category discrimination[5], and normalization[6]. However, the results are not as sound as they predicted. Cho and Wüthrich (1999) found that the most important factor influencing prediction accuracy is the selection of the data sources. Jaybhay et al (2012) listed four forecasting methods based on textual mining: 1. Technical analysis methods, 2.

[4] The simple term frequency is based on counting the words "up"," down", and "steady" on web pages on the Hang Seng Indexes.

[5] Category discrimination is a method for dividing different texts on the Hang Seng Indexes for further analysis.

[6] Normalization is the calculation method for obtaining each day's maximum value based on the results of category discrimination.

fundamental analysis techniques, 3. traditional time series prediction, and 4. machine learning methods.

Derakhshan and Beigy (2019) introduced the LDA-based method7 and LDA-POS method8 to the analysis. They sought to avoid problems in the text. Social media text is usually short, and it contains many misspellings, uncommon grammar constructions, and other issue. They reconstructed the text based on the topic label and distribution, which achieves some positive results for specific text data; however, for other data, this method did not work. This paper highlights a phenomenon model that needs to be adjusted via unstructured data.

Das and Chen (2007) built an analysis system based on the statistical dictionary method9; furthermore, they listed five kinds of clusters for message interpretation, and the decision was positive, negative, or neutral. The five classifiers of the naive classifier10, vector distance classifier 11 , adjective-adverb phrase classifier 12 , and Bayesian classifier13 have different structures based on different languages and grammar.

Schumaker et al (2012) used machine learning in sentiment analysis. They found that this method was best able to predict subjective articles in directional accuracy (positive or negative tendency of price from text data) and trading returns but not closeness. They also found a phenomenon different from common sense: negative sentiments should be indicative of downward price movement. Oliveira et al (2013) used two error metrics (the MAPE14 and RMSE15) to measure the results of models. In their paper, Guo et al (2017)

[7] The LDA-based method is an affective computing method based on the generative probability of the mixture of latent topics, and each topic is a probability distribution.

[8] The LDA-POS method, which is an optimized version of the LDA-based method, is an affective computing method that incorporates speech tags into topic modeling, which is an optimized method comparing with LDA-based method.

[9] The statistical dictionary method is an affective computing method. The text data are classified into three types by statistics and an emotion dictionary: bullish, neutral and bearish.

[10] The naive classifier is an algorithm based on the counting of words with positive and negative connotations.

[11] The vector distance classifier is an algorithm in which hand-tagged text data are calculated in vector function for classifying text data.

[12] The adjective-adverb phrase classifier is an algorithm in which the word count process is based on adjectives and adverbs from text data.

[13] The Bayesian classifier relies on a multivariate application classification algorithm that is based on Bayes' theorem.

[14] Mean Absolute Percentage Error

[15] Root Mean Square Error

showed that sentiment data do not lead stock prices all the time, and a lead-lag structure appears. They also introduced the thermal optimal path method16 for their time series analysis.

In this thesis, the idea of an affective computing algorithm is combined with the advantages of the statistical dictionary method and term frequency. For further analysis, the lag influence between time series is analyzed by the TLCC algorithm, and the results of the model are evaluated using a series of error metrics and standards from the VAR model.

[16] The Thermal Optimal Path method is an algorithm for identifying and quantifying the lead-lag structure in two different time series data.

### *2.2. Behavioural Finance and Efficient Market Theory*

The basis of this article is whether internet public opinion can affect the direction of the stock market or whether the two factors have a mutual influencing relationship. To realize research and judgment of this foundation, behavioral finance and efficient market theory are introduced here.

Traditional financial theory believes that people's decision making in the market is based on harsh assumptions, such as rational expectations, risk aversion, utility maximization, and discretion. Fama (1965) proposed an efficient market hypothesis that is based on ideal market conditions. Under the assumption of rational people, stock prices reflect the balance between supply and demand, and arbitrage behavior makes stock prices move rapidly enough to make the two factors equal. A stock price fully reflects all the information related to the asset. However, in reality, the information is not always effective. Not every market participant, especially retail investors, can act completely rationally and procedurally according to theory and probability, such as adopting a quantitative investment algorithm. The irrational behavior of market participants plays a huge role in changes in financial markets. In his book on behavioral finance, Ritter (2003), states that behavioral finance encompasses research that drops the traditional assumptions of expected utility maximization with rational investors in efficient markets. The two building blocks of behavioral finance are cognitive psychology (how people think) and the limits to arbitrage (when markets will be inefficient).

Research on behavioral finance has further challenged the prerequisites of the efficient market theory. Market participants have widespread irrational behaviors; arbitrage trading has various limitations in the real world, which cannot achieve the expected effects in theory; and transactions often appear out of time.

Behavioral finance mainly focuses on the following directions: the investment behavior of individual investors, especially retail investors; the mutual influence between investors and the result of this influence in the medium and long term; the herd effect; the small company effect; equity premiums; and various extreme behaviors in specific turbulent times.

### *2.3. Chinese Stock Market*

For historical and political reasons, the Chinese stock market includes the Shanghai Stock Exchange and Shenzhen Stock Exchange, and they have several unique features. (i) Fernald and Rogers (2002) stated that the government control of Chinese capital makes it difficult for retail investors to invest outside of China. The A shares17 and B shares18 systems were developed to conduct capital control. (ii) Yao et al (2014) revealed the reality that domestic retail investors dominate A-share markets while foreign institutional traders mainly dominate B-share markets. (iii) Kong and Wang (2014) found that in the Chinese capital markets, (1) order-based manipulation from the government affects the liquidity and trading behavior of the Chinese capital markets and (2) the manipulator pretends to be informed or expects to be seen as informed by choosing the "right" time to implement the manipulation. (iv) Chong et al (2017) found that recommendations from analysts, the short-term horizons of retail investors, and risk aversion are the main reasons for herding behavior in the Chinese stock markets. (v) The thesis of Hung (2009) found that the weak-form efficient market hypothesis was rejected for both the Shanghai Stock Exchange and Shenzhen Stock Exchange. With the ongoing progress of deregulation and liberalization, the efficiency of the Chinese stock market also gradually improves. (vi) Ni et al (2015) performed a nonlinear effect test of the investor sentiment and returns in the Chinese stock market using a panel quantile regression model. They found that in the Chinese stock market, investors have a notable cognitive bias and speculation tendency. (vii) Changsheng and Yongfeng (2012) researched the Chinese stock market and found that investor sentiment had a significant explanatory ability for both retail investors' favorable stocks and value stocks, which shows that when investors are bullish (bearish), these stocks will generate higher (lower) excess returns. It is important to note that investor sentiment is an important systemic risk factor in asset pricing models. In conclusion, the Chinese stock market is an immature stock market that is mainly composed of retail investors and is strongly regulated and influenced by the Chinese government with restricted capital controls.

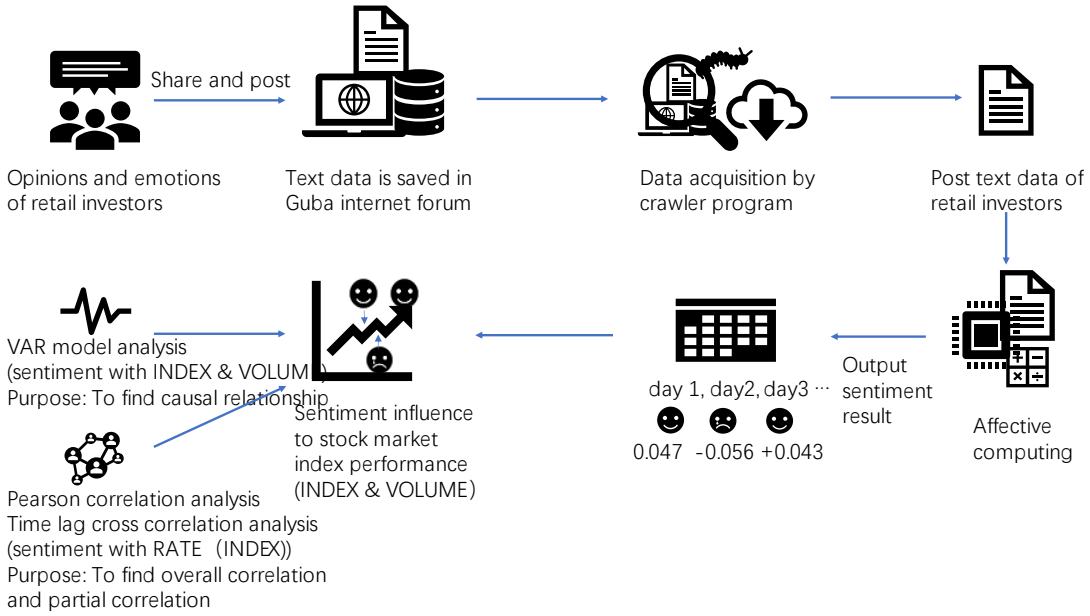
[17] A shares are issued in China, with RMB as the denominated currency, for individuals, organizations, and companies in China to subscribe and trade.

[18] B shares are issued in China, with RMB as the denominated currency, for overseas individuals, organizations and companies to subscribe and trade.

## 3.Affective computing and model analysis

The general idea of affective computing and model analysis in this thesis is as follows.

FIGURE 2 – Technical route.




### *3.1 Data collection*

From the literature search, it can be concluded that the following are the main types of text-based sentiment analyses in the Chinese market. The first type is based on public media, such as portal websites, newspapers, and periodicals. The second type is based on various indicator data in the transaction process. The third type is based on stock-based internet forums, where stockholders visit, share and comment on their feelings and opinions. The fourth type is based on Weibo, which is equivalent to Twitter in China. Most followers are strangers, and a few followers are acquaintances. The fifth type is social applications for acquaintances, such as WeChat, which is equivalent to WhatsApp in China.

In the Chinese market, there is a phenomenon of media control that cannot be ignored. Almost all public media or public social media comply with content censorship, and these media should comply with the regulations and intentions of the government. Relatively large institutions and opinion leaders have more opportunities to monopolize public influence. This type of text data cannot be used to represent the opinions of retail investors.

The second type is based on various data indicators in the transaction process that reflect emotions, such as the stock turnover rate. There is a certain deviation between the emotions and specific behaviors of market participants. This kind of deviation includes the deviation over time and the deviation in actions.

The third type is stock forums, where stockholders gather together online. Stockholders freely express their opinions and feelings to each other. The government restrictions on stock forums are much looser than those on portal websites, newspapers, and periodicals. Shareholders also comment on each other's posts. These types of text data are instant, real, and abundant for analysis.

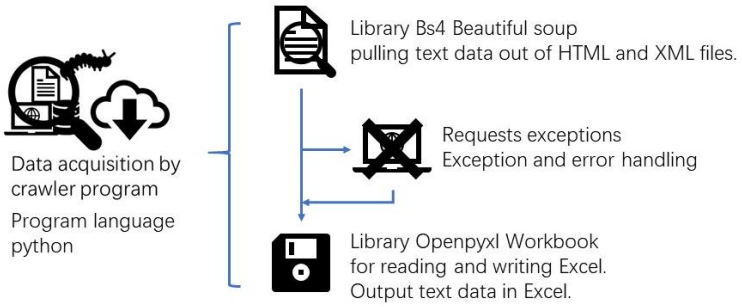
The fourth and fifth types are social applications between partial acquaintances or full acquaintances, respectively. Social applications such as Weibo and WeChat are mostly used to share users' daily lives. There is a cultural taboo in China that people showing one's wealth directly on social media is regarded as rude and not well educated. Personal information about buying and selling stocks normally is not posted on social media. Only in the era of a big bull market or big bear market do people share their transaction information and feelings. Text data that are effective for analysis are difficult to collect.

To conclude, in this thesis, the third type of text data from retail investor forums were chosen for analysis. The data of the Shanghai Securities Composite Index and the total trading volume of the Shanghai Stock Exchange are from the official website of the Shanghai Stock Exchange.

### *3.2Text data acquisition by crawler*

To automatically acquire big text data from the target internet forum of retail investors, a crawler was programmed as follows.

FIGURE 3 – Structure of the Crawler.




The largest stock internet forum in China is https://guba.eastmoney.com/. The forum is divided into several different sections. There are separate discussion areas for various investment targets. In this forum, stockholders express their opinions and discuss and comment all day. After years of development, this forum has become the most active platform for Chinese investors. In this thesis, the Shanghai Securities Composite Index investor forum is used as the data source.

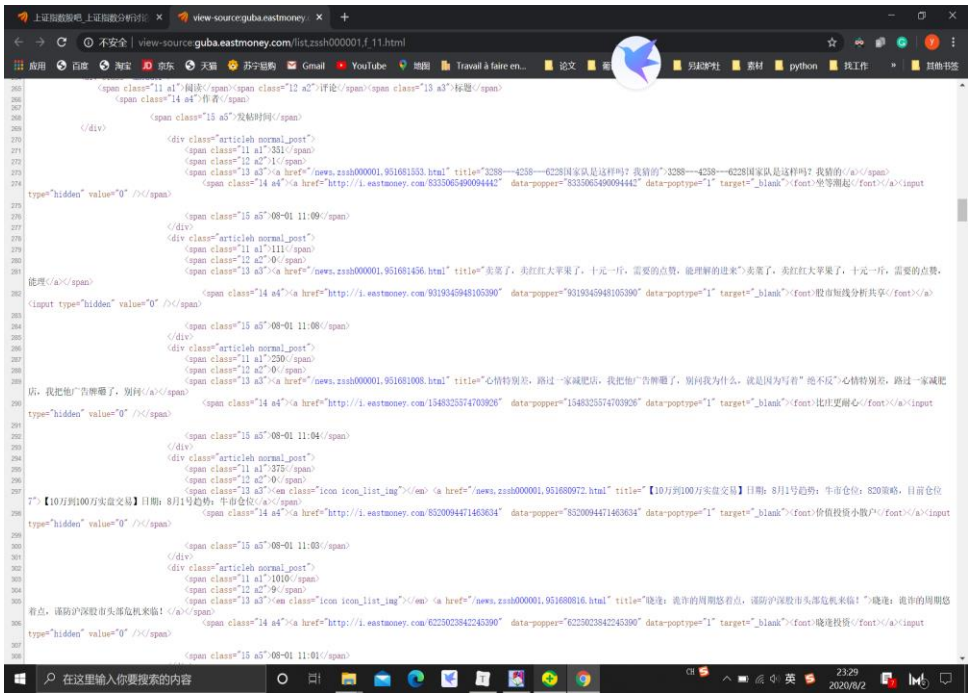
FIGURE 4 – Guba Shanghai Securities Composite Index Forum.


The Shanghai Securities Composite Index is composed of all the public listed stocks on the Shanghai Stock Exchange, including A shares (issued in China, with RMB as the denominated currency, for individuals, organizations, and companies in China to

subscribe and trade) and B shares (issued in China, with RMB as the denominated currency for overseas individuals, organizations and companies to subscribe and trade). The Shanghai Securities Composite Index reflects the change in the price level of all the listed stocks on the Shanghai Stock Exchange. Finally, the closing price data of each trading day and the total trading volume of each trading day of the Shanghai Securities Composite Index are used for analysis.

Each retail investor's post is composed of several components. The components include the title, date, user name, body content, and other investors' replies. The title and date of each post are used as data.

FIGURE 5 – Guba Shanghai Securities Composite Index forum source code.



By observing the source code, the required data are encapsulated in a specific key-value pair. To automatically obtain data, the crawler method was used. Python was used as the program language.

FIGURE 6 – Library called by the crawler.

```
from bs4 import BeautifulSoup
from openpyxl import Workbook
from requests.exceptions import RequestException
import re
import time
```

One of the Python libraries used was Bs4 Beautiful Soup: Richardson (2007) developed Beautiful Soup, which is a Python crawler library for extracting data from HTML and XML files. The library works with parsers to provide idiomatic ways of navigating, searching, and modifying parse trees.

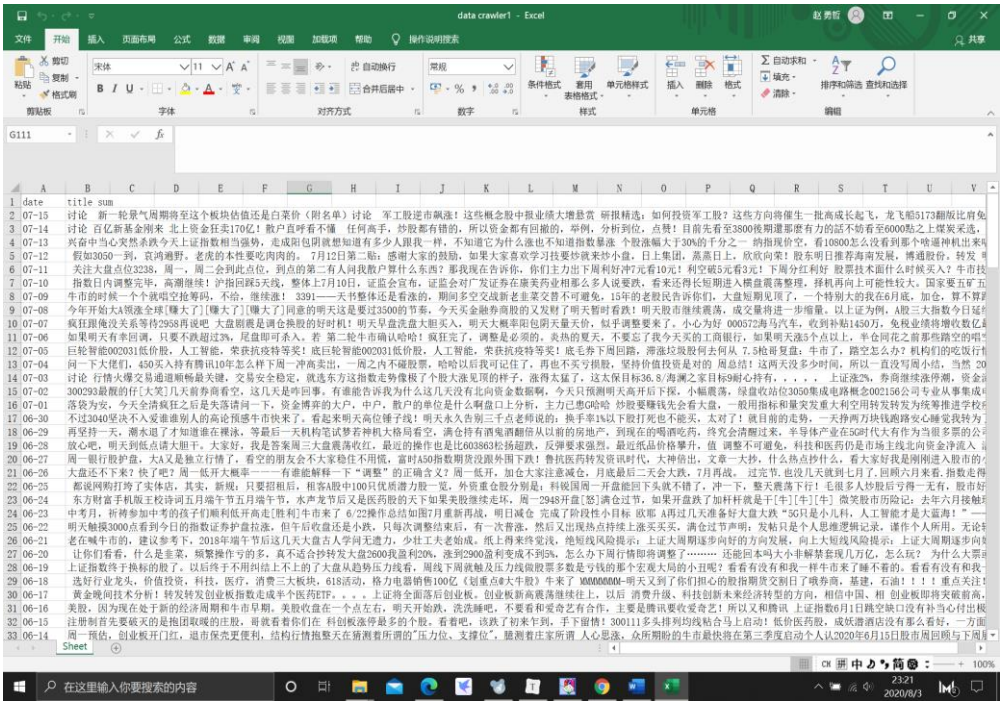
Another Python library used was openpyxl Workbook: Gazoni and Clark (2018) developed openpyxl as an open source project, and the openpyxl module is a Python library for reading and writing Excel 2010 documents.

Another Python library used was request.exceptions: Reitz (2020) developed this library to handle exceptions and errors during connections.

The crawler automatically extracted the titles of all the retail posts that were published from January 1, 2020, to July 10, 2020. Then, the algorithm was used to summarize all the titles posted each day. The output data are composed of the corresponding date and sum of all the titles. The reason why only the titles are used is that after text analysis, retail investors summarize their posts using titles, and the title has been shown to have sufficient sentiment tendencies. The content of the main body of the posts published by retail investors is relatively random and is often mixed with various content that is not related to the subject, which makes various errors more likely to occur in subsequent text analysis.

After obtaining the data, in the data sorting process, data from nontrading days were deleted. Furthermore, the data of exception strings that could not be processed were also deleted.

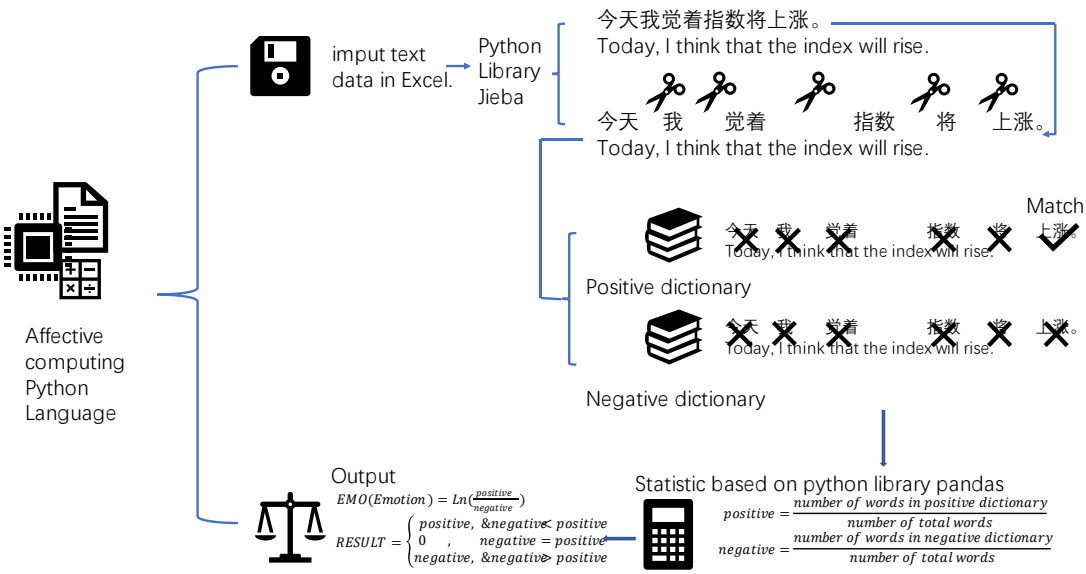
FIGURE 7 – Text data acquired by the crawler.



### *3.3 Sentiment dictionary preparation*

The main technical realization path of this paper is affective computing, which is a dictionary sentiment calculation. The structure of affective computing is as follows.

FIGURE 8 – The structure of affective computing.




First, a proper dictionary for emotional calculation is essential. The main technical realization path of this paper is the dictionary sentiment calculation method. First, a proper dictionary for emotional calculation is essential. At present, most Chinese emotional dictionaries are general purpose dictionaries, which are mainly composed of adjectives and adverbs. Zhu et al (2006) and Chen et al (2018) developed sentiment dictionaries such as HOWNET and NTUSD, respectively. These dictionaries are not professional financial dictionaries. The text data for research on business issues should not include noncommercial data sets, as dictionaries constructed in this way will cause larger deviations in the results. The sentiment dictionaries HOWNET and NTUSD were used as the initial dictionaries. The commonly used vocabulary of this stock forum and the common oral vocabulary of retail investors in China were collected. Finally, the Chinese financial sentiment dictionary dedicated to the Guba Shanghai Securities Composite Index forum was constructed.

TABLE 1

EXAMPLE OF POSITIVE WORDS AND NEGATIVE WORDS DICTIONARY

| | |
|---|---|
| positive | |
| 涨 | rise |
| 赚 | gain profit |
| 猛 | rush to gain profit |
| 进 | buy |
| 冲 | rush to buy |
| 上 | price will go up |
| 长多 | buy |
| 护盘 | Government or institution investor protect market |
| 利多 | information in market which shows it is better to buy |
| 牛市 | bull market |
| ... | |
| | |
| negative | |
| 跌 | price of stock goes down |
| 亏 | loss money |
| 走 | sell |
| 怂 | scared to buy |
| 下 | price go down |
| 拨档 | When investors are long, the stock price falls, and the stock price is expected to continue to fall. |
| 崩 | stock market is out of control to go down. |
| 利空 | information which leads market go down. |
| 暴雷 | suddenly the bad news come out to public. |
| 熊市 | bear market |

### *3.4 The jieba method for autodividing sentence*

Another Python library used was Jieba: In Junyi (2015), the Jieba method was used to cut sentences into words based on Chinese language grammar and tradition. The Jieba method is composed of three modes. (i) Accurate mode: The content of the text is accurately divided according to the grammatical structure and daily usage habits, and there is no redundant separate vocabulary. (ii) Complete mode: All possible words are scanned in segments, and there is single character redundancy.

(iii) Search engine mode: Under the guidance of the precise mode, long words are also segmented from inside. The precise mode segmentation method is used for analysis in this thesis, and the function used is jieba.lcut (content).

### *3.5 Construction and result of the sentiment function*

Then, the Python library Pandas was used for statistics. The divided words were entered in the positive dictionary and the negative dictionary one by one for comparison and counting.

The Python library Pandas was also used: McKinney (2011) developed Pandas, a Python library for big data structures. Pandas has been widely used in statistics, finance, social sciences, and many other fields. The library includes functions of integrated and intuitive routines for performing common data control and analysis. Pandas can be used for data collection, data analysis, and data cleaning purposes. The output of each day's sentiment value is based on the percentage of the statistical value of all the positive and negative sentiment words to the total number of words from each text data. Positive and negative algorithms combine the advantages of the statistical dictionary method and term frequency.

$$(1)\ positive = \frac{number\ of\ words\ in\ positive\ dictionary}{number\ of\ total\ words}$$

$$(2)\ negative = \frac{number\ of\ words\ in\ negative\ dictionary}{number\ of\ total\ words}$$

*Source: stake overflow.com (2020)*

The counted positive and negative words are nouns, adjectives, adverbs and grammar instructed slang. The output is a list of positive sentiment values and a list of negative sentiment values. To better use positive data and negative data for analysis, the emo (emotion) function is constructed for analysis. The following result function is constructed for the direction analysis of the Shanghai Securities Composite Index.

$$(3)\ EMO(Emotion) = Ln(\frac{positive}{negative})$$

$$(4)\ RESULT = \begin{cases} positive, & negative < positive \\ 0 \quad , & negative = positive \\ negative, & negative > positive \end{cases}$$

Figure 9 – Positive sentiment result.

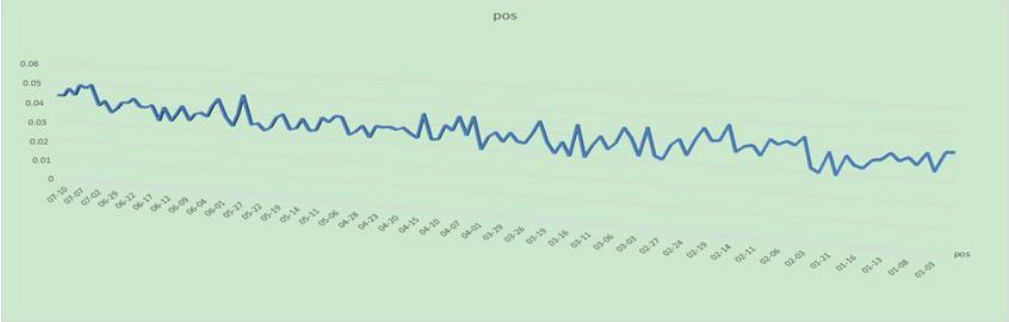



Figure 10 – Negative sentiment result.

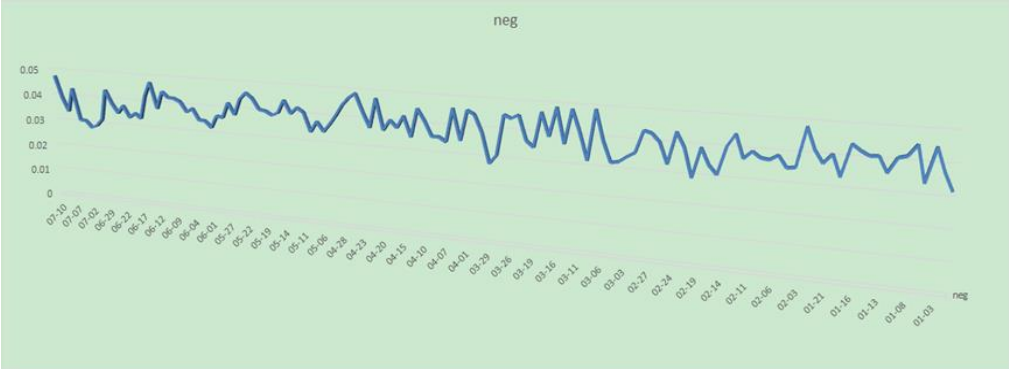
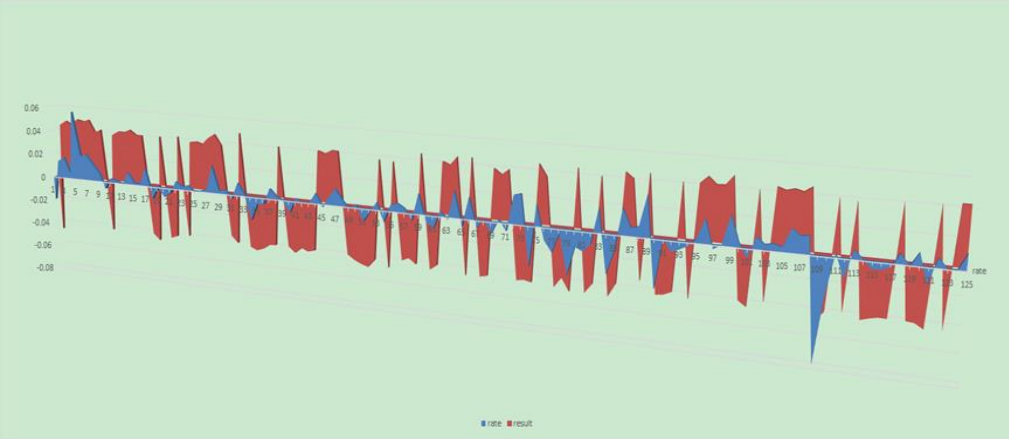



Figure 11 – Affective computing result（RESULT）and Shanghai Securities Composite Index change rate(RATE).



It can be intuitively seen from the image that sentiment has a certain consistency with the rise and fall of the Shanghai Securities Composite Index. In addition, there are still some lag differences in specific sections.

### *3.5 VAR model analysis (SENTIMENT & INDEX)*

Next, the VAR model was constructed to explore the relationship between the sentiment of retail investors and the Shanghai Securities Composite Index. Qin (2011) introduced the history of the VAR model. The VAR approach arises from a fusion of the Cowles commission tradition and time series statistical methods, and it is catalyzed by the rational expectations (RE) movement.

First, to find the relationship between the variable sentiment value (EMO) of retail investors, the unit root of the Shanghai Securities Composite Index (INDEX) and the first-order difference of the Shanghai Securities Composite Index needed to be tested. The first-order difference in the Shanghai Securities Composite Index represents the change in the Shanghai Securities Composite Index. The test results were as follows. The results showed that no root was outside the unit circle, and the model satisfied the stability condition.

Figure 12 – EMO INDEX unit root test.

Roots of Characteristic Polynomial
Endogenous variables: EMO INDEX
Exogenous variables: C
Lag specification: 1 1
Date: 08/09/20 Time: 18:36

| Root | Modulus |
|---|---|
| 0.978915 | 0.978915 |
| 0.149239 | 0.149239 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

Figure 13 – EMO D(INDEX) unit root test.

Roots of Characteristic Polynomial
Endogenous variables: EMO D(INDEX)
Exogenous variables: C
Lag specification: 1 1
Date: 08/15/20 Time: 14:37

| Root | Modulus |
|---|---|
| 0.163206 | 0.163206 |
| 0.025915 | 0.025915 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

After confirming that the unit root test was passed, the lag order of the VAR model was determined by the VAR lag order selection criteria.

Figure 14 – EMO INDEX VAR Lag order selection criteria.

VAR Lag Order Selection Criteria
Endogenous variables: EMO INDEX
Exogenous variables: C
Date: 08/09/20 Time: 18:27
Sample: 1 125
Included observations: 117

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -730.0239 | NA | 932.1339 | 12.51323 | 12.56045 | 12.53240 |
| 1 | -563.8886 | 323.7509* | 58.32053* | 9.741686* | 9.883336* | 9.799194* |
| 2 | -561.9955 | 3.624381 | 60.46420 | 9.777701 | 10.01378 | 9.873548 |
| 3 | -559.3457 | 4.982654 | 61.88716 | 9.800781 | 10.13130 | 9.934966 |
| 4 | -558.1450 | 2.216584 | 64.94155 | 9.848633 | 10.27358 | 10.02116 |
| 5 | -554.8395 | 5.989428 | 65.75033 | 9.860505 | 10.37989 | 10.07137 |
| 6 | -552.0466 | 4.965142 | 67.17006 | 9.881139 | 10.49496 | 10.13034 |
| 7 | -550.1719 | 3.268708 | 69.72456 | 9.917469 | 10.62572 | 10.20501 |
| 8 | -547.2788 | 4.945490 | 71.14917 | 9.936390 | 10.73907 | 10.26227 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Figure 15 – EMO INDEX VAR lag 1 model.

Vector Autoregression Estimates
Date: 08/09/20 Time: 18:31
Sample (adjusted): 2 125
Included observations: 124 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | EMO | INDEX |
|---|---|---|
| EMO(-1) | 0.151329 | 21.07640 |
| | (0.09438) | (18.4605) |
| | [ 1.60341] | [ 1.14170] |
| INDEX(-1) | 8.21E-05 | 0.976825 |
| | (0.00016) | (0.03074) |
| | [ 0.52220] | [ 31.7808] |
| C | -0.242087 | 70.19902 |
| | (0.46025) | (90.0257) |
| | [-0.52599] | [ 0.77977] |
| R-squared | 0.030699 | 0.905982 |
| Adj. R-squared | 0.014677 | 0.904428 |
| Sum sq. resids | 6.074675 | 232412.7 |
| S.E. equation | 0.224062 | 43.82655 |
| F-statistic | 1.916110 | 582.9958 |
| Log likelihood | 11.05311 | -643.1797 |
| Akaike AIC | -0.129889 | 10.42225 |
| Schwarz SC | -0.061656 | 10.49049 |
| Mean dependent | -0.001774 | 2928.811 |
| S.D. dependent | 0.225725 | 141.7661 |
| Determinant resid covariance (dof adj.) | | 52.78554 |
| Determinant resid covariance | | 50.26229 |
| Log likelihood | | -594.7666 |
| Akaike information criterion | | 9.689784 |
| Schwarz criterion | | 9.826249 |

The lag 1 is significant for all kinds of information criteria, and it was selected to build a model for comparison. In the table of the VAR lag 1 model, the t-statistic is 1.14170, which is not significant at the 5% significance level.

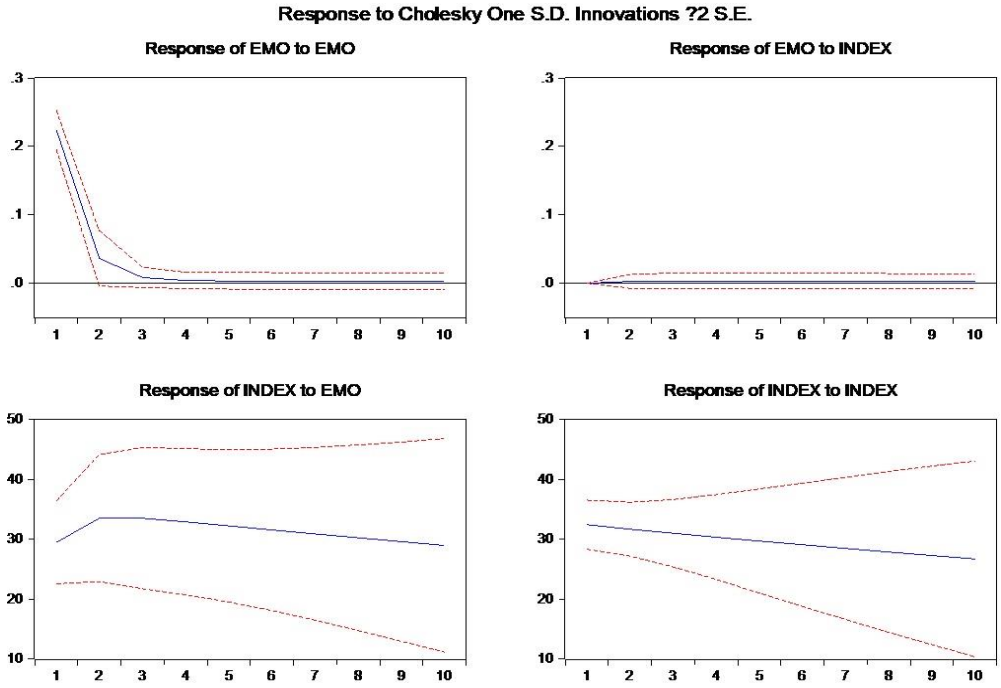
Figure 16 – EMO INDEX Pairwise Granger causality tests..

Pairwise Granger Causality Tests
Date: 08/09/20 Time: 18:23
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| INDEX does not Granger Cause EMO | 123 | 0.00388 | 0.9961 |
| EMO does not Granger Cause INDEX | | 2.11567 | 0.1251 |

Through Pairwise Granger Causality Test, it can be found that both null hypothesis INDEX dose not Granger Cause EMO and EMO does not Granger Cause INDEX.

Figure 17 – EMO INDEX impulse analysis.




It can be seen from the figure that EMO has a positive impact on the subsequent phase of EMO. After the second phase, the impact gradually disappears. This finding shows that the influence of the sentiment of retail investors reaches the strongest phase one day later and gradually disappears in the next two days.

EMO's response to INDEX is not obvious, indicating that under this model, INDEX has no obvious direct influence on the sentiment of retail investors.

In contrast, the response of INDEX to EMO is significant, reaching its maximum influence in the second phase and then gradually falling. This finding shows that it takes

two days for investors' emotions to reach the maximum impact and then gradually decrease. Furthermore, the impact of the index on itself is positive and gradually decreases.

To further analyze the relationship between investor sentiment and the change in the Shanghai Securities Composite Index, a VAR model was established to analyze the relationship between the first-order difference of EMO and INDEX.

Figure 18 – EMO D（INDEX）VAR Lag order selection criteria

VAR Lag Order Selection Criteria
Endogenous variables: EMO D(INDEX)
Exogenous variables: C
Date: 08/10/20 Time: 13:21
Sample: 1 125
Included observations: 116

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -563.8150 | NA* | 59.12637* | 9.755430* | 9.802906* | 9.774703* |
| 1 | -561.4612 | 4.585717 | 60.83000 | 9.783814 | 9.926241 | 9.841631 |
| 2 | -559.6851 | 3.399174 | 63.21295 | 9.822157 | 10.05953 | 9.918519 |
| 3 | -558.0788 | 3.018714 | 65.88870 | 9.863427 | 10.19576 | 9.998334 |
| 4 | -554.6063 | 6.406160 | 66.51271 | 9.872522 | 10.29980 | 10.04597 |
| 5 | -552.1697 | 4.411091 | 68.36537 | 9.899477 | 10.42171 | 10.11147 |
| 6 | -550.8645 | 2.317839 | 71.67030 | 9.945940 | 10.56312 | 10.19648 |
| 7 | -549.1672 | 2.955591 | 74.64921 | 9.985642 | 10.69778 | 10.27473 |
| 8 | -546.3639 | 4.784970 | 76.30755 | 10.00627 | 10.81336 | 10.33391 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

.

Figure 19 – EMO D（INDEX）lag 1 Pairwise Granger causality tests.

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:31
Sample: 1 125
Lags: 1

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| D(INDEX) does not Granger Cause EMO | 123 | 0.00174 | 0.9668 |
| EMO does not Granger Cause D(INDEX) | | 0.30471 | 0.5820 |

Figure 20 – EMO D（INDEX）lag 2 Pairwise Granger causality tests.

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:29
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| D(INDEX) does not Granger Cause EMO | 122 | 0.22761 | 0.7968 |
| EMO does not Granger Cause D(INDEX) | | 0.95721 | 0.3869 |

Through lag order analysis, only lag 0 is significant. It is impossible to establish a suitable VAR model for analysis. In addition, in the Pairwise Granger Causality Tests of lag1 and lag2, it was found that the null hypothesis could not be rejected. Therefore, it is impossible to determine that investor sentiment is the reason for leading the change of the Shanghai Securities Composite Index.

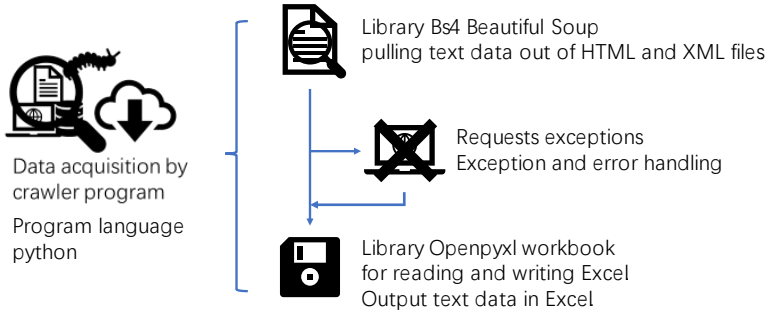
### *3.6 Pearson correlation and TLCC analysis (SENTIMENT & INDEX)*

There is no causal relationship between the sentiment of retail investors and the Shanghai Securities Composite Index that can be proven by previous VAR models. In addition, there is no causal relationship between the sentiment of retail investors and the first-order difference of the Shanghai Securities Composite Index that can be proven by previous VAR models. Although causal relationships fail to be proven, there might be an overall correlation or partial correlation between the sentiment of retail investors and the change rate of the Shanghai Securities Composite Index. To avoid the problem that the numbers of indexes are much larger than the sentiment result, which makes the correlation calculation inaccurate, a rate function was built.

$$(5)\ RATE(INDEX) = \frac{price\ of\ index - price\ of\ index\ (a\ day\ before)}{price\ of\ index\ (a\ day\ before)}$$

Next, the Pearson method and the TLCC method were used to find the correlation relationship between two time series.

Figure 21 – Structures of the Pearson correlation and time lag cross correlation analyses




To quantify the relationship between two time series, some libraries based on Python were used, such as Pandas, NumPy, matplotlib, seaborn and SciPy.

Python library NumPy: Oliphant (2006) built NumPy on a successful numeric array object. The goal is to create the cornerstone of a useful environment for scientific computing. A large number of dimensional multidimensional matrix calculations are supported by NumPy, and it can also provide a large number of mathematical function libraries for array operations. Pandas is based on NumPy. High-performance matrix calculation support can be provided by NumPy.

Python library matplotlib: Hunter and Dale (2007) The matplotlib is a library for making 2D plots of arrays in python

Python library seaborn: Sheppard (2012) developed seaborn to provide a number of advanced data visualized plots and a general improvement in the default appearance of matplotlib-produced plots.

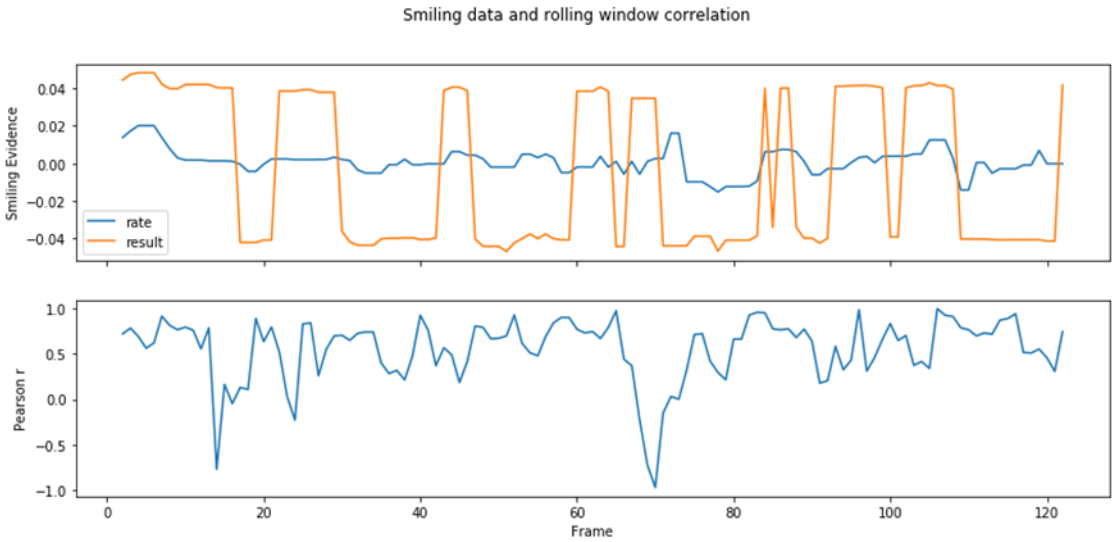
Python library SciPy: Oliphant (2004) developed SciPy as a collection of mathematical algorithms and convenience functions that is built on the numerical extension of Python. It can be used for the manipulation and visualization of data.

The Pearson correlation coefficient statistical guide (2020) states that the Pearson correlation is a measure of the linear correlation between two variables and how the two variables change together over time. The correlation coefficient can be used to reveal the degree of correlation (0.5 (-0.5)-1.0 (0.5) means a strong correlation, 0.3 (-0.3)-0.5 (-0.5) means a medium correlation, and 0.1 (-0.1)-0.3 (-0.3) means a weak correlation). When

the correlation coefficient r is greater than 0 and less than 1, it indicates a positive correlation between x and y. When r is greater than -1 and less than 0, it indicates a negative correlation between x and y. It is unusual to regard the Pearson correlation as a measure of full-time synchronization. Therefore, the Pearson correlation cannot be used to judge the directionality between two variables, and it cannot distinguish which variable plays a leading role and which variable is just following.

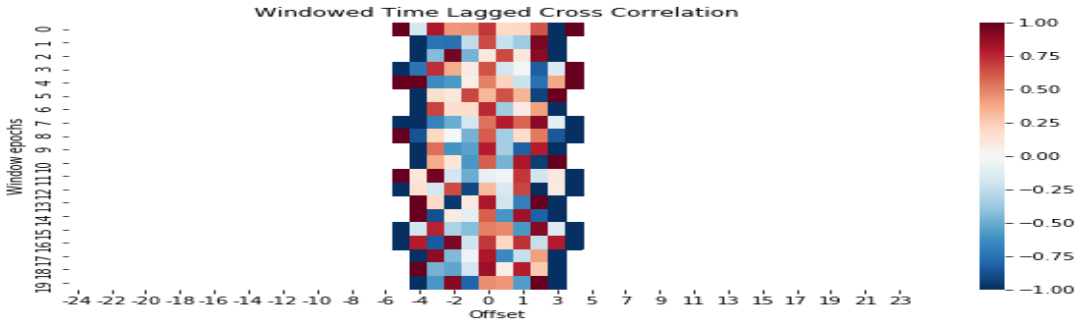
After calculation, the Pearson correlation coefficient was 0.5159316953442752 and the p-value was 7.37838628287929e-10. The results show that the sentiment of retail investors and the change rate of the Shanghai Securities Composite Index are strongly correlated.

Figure 22 – Pearson analysis.




In addition, to measure the local synchronization, the sliding window method was adopted, and the Pearson correlation was repeatedly calculated in all the sliding windows until all the variables were covered by the window. The figure above shows the result of the synchronization calculation at each moment. It appears that for the majority of the time, there is positive synchronization between two time series.

Figure 23 – TLCC analysis.




The literature search shows that some scholars find that there is a lag difference between the sentiment of retail investors and changes in the stock market. To further explore the influencing relationship between the two variables, the time lag cross-correlation of different windows was calculated for analysis. The time lag cross-correlation was repeatedly calculated in multiple time windows. By comparing the differences in the scores in the interaction, the initiator and the followers was found. The time series was divided into 20 equal size time windows, and then, the cross-correlation of each time window was calculated. Through analysis of the results, it was found that there is a lag in the mutual influence relationship between the sentiment of retail investors and the change rate of the Shanghai Securities Composite Index.

### *3.7 VAR model analysis (SENTIMENT & VOLUME)*

To find the relationship between the variable sentiment value (EMO) and the trading volume of the Shanghai Securities Composite Index, the unit root should be tested. Because the real trading volume is normally a huge number, the logarithms of those numbers have been taken. The test results are as follows.

Figure 24 – EMO VOLUME unit root test.

Roots of Characteristic Polynomial
Endogenous variables: EMO VOLUME
Exogenous variables: C
Lag specification: 1 2
Date: 08/10/20 Time: 13:45

| Root | Modulus |
|---|---|
| 0.975540 | 0.975540 |
| -0.480933 | 0.480933 |
| 0.348395 | 0.348395 |
| -0.173426 | 0.173426 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

The results are all within the unit circle, and the data are stationary and can be used for modeling. The result shows that no root lies outside the unit circle, and the two VARs satisfy the stability conditions.

Figure 25 – EMO VOLUME var lag order selection criteria.

VAR Lag Order Selection Criteria
Endogenous variables: EMO VOLUME
Exogenous variables: C
Date: 08/10/20 Time: 13:46
Sample: 1 125
Included observations: 117

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -6.330924 | NA | 0.003953 | 0.142409 | 0.189626 | 0.161578 |
| 1 | 99.54441 | 206.3212 | 0.000693 | -1.599050 | -1.457400 | -1.541542 |
| 2 | 110.2646 | 20.52413* | 0.000618* | -1.713925* | -1.477842* | -1.618078* |
| 3 | 114.0927 | 7.198049 | 0.000620 | -1.710986 | -1.380469 | -1.576800 |
| 4 | 115.8577 | 3.258508 | 0.000644 | -1.672781 | -1.247831 | -1.500257 |
| 5 | 115.8984 | 0.073844 | 0.000689 | -1.605101 | -1.085718 | -1.394238 |
| 6 | 119.7197 | 6.793370 | 0.000692 | -1.602046 | -0.988230 | -1.352844 |
| 7 | 124.6673 | 8.626595 | 0.000682 | -1.618245 | -0.909995 | -1.330704 |
| 8 | 129.0757 | 7.535727 | 0.000678 | -1.625226 | -0.822543 | -1.299346 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

The testing the lag order selection criteria found that lag 2 is significant.

Vector Autoregression Estimates
Date: 08/10/20 Time: 13:42
Sample (adjusted): 3 125
Included observations: 123 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | EMO | VOLUME |
|---|---|---|
| EMO(-1) | 0.081984 | 0.153052 |
| | (0.09790) | (0.05100) |
| | [ 0.83742] | [ 3.00078] |
| EMO(-2) | 0.088792 | -0.026536 |
| | (0.09640) | (0.05022) |
| | [ 0.92104] | [-0.52834] |
| VOLUME(-1) | 0.245141 | 0.587592 |
| | (0.17562) | (0.09149) |
| | [ 1.39586] | [ 6.42221] |
| VOLUME(-2) | -0.152757 | 0.364912 |
| | (0.17460) | (0.09096) |
| | [-0.87489] | [ 4.01163] |
| C | -2.446018 | 1.265450 |
| | (2.29063) | (1.19336) |
| | [-1.06783] | [ 1.06041] |
| R-squared | 0.063711 | 0.840198 |
| Adj. R-squared | 0.031972 | 0.834781 |
| Sum sq. resids | 5.851959 | 1.588312 |
| S.E. equation | 0.222695 | 0.116018 |
| F-statistic | 2.007353 | 155.1032 |
| Log likelihood | 12.76315 | 92.96559 |
| Akaike AIC | -0.126230 | -1.430335 |
| Schwarz SC | -0.011914 | -1.316018 |
| Mean dependent | -0.002827 | 26.44295 |
| S.D. dependent | 0.226342 | 0.285428 |

| | |
|---|---|
| Determinant resid covariance (dof adj.) | 0.000555 |
| Determinant resid covariance | 0.000510 |
| Log likelihood | 117.1328 |
| Akaike information criterion | -1.741996 |
| Schwarz criterion | -1.513364 |

Figure 26 – EMO VOLUME VAR model.

Figure 27 – EMO VOLUME VAR model.

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:39
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| VOLUME does not Granger Cause EMO | 123 | 1.24079 | 0.2929 |
| EMO does not Granger Cause VOLUME | | 4.58500 | 0.0121 |

In vector autoregression estimates table, the t-statistics of EMO (-1), VOLUME (-1) and VOLUME (-2) are all significant. The r-squared is 0.840198, which is quite high. Through the pairwise Granger causality test, the p-value is 0.0121, which is significant. The null hypothesis (EMO does not Granger cause VOLUME) can be rejected. The sentiment of retail investors is the reason for the change in trading volume.

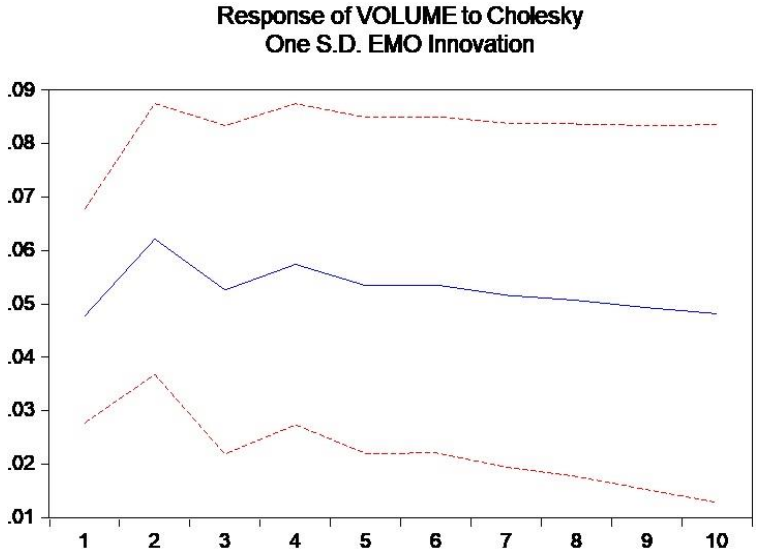
VAR Model:

VOLUME = C(2,1)*EMO(-1) + C(2,2)*EMO(-2) + C(2,3)*VOLUME(-1) + C(2,4)*VOLUME(-2) + C(2,5)

VAR Model - Substituted Coefficients:

VOLUME = 0.1530517258*EMO(-1) - 0.0265355078095*EMO(-2) + 0.587591946226*VOLUME(-1) + 0.364911869886*VOLUME(-2) + 1.26544984901

Figure 28 – EMO VOLUME impulse analysis




The results of the impulse response analysis show that the influence of investor sentiment on trading volume gradually increases from the first period to the second period, the influence decreases in the third period, and the influence increases in the fourth period. Then, the influence gradually diminishes.

# 4.Conclusions and future research

## *4.1 Conclusions*

By analyzing the relationship between the sentiment of retail investors and the Shanghai Securities Composite Index and the relationship between the sentiment of retail investors and the total trading volume of the Shanghai Stock Exchange, the results remind retail investors to improve their rational investment awareness and provide a reference for the Chinese government to improve the quality of market regulation to protect retail investors. In the VAR model (EMO INDEX), the pairwise Granger causality test found that neither null hypothesis was supported: INDEX does not Granger cause EMO and EMO does not Granger cause INDEX. The results of the EMO INDEX impulse analysis show that the response of INDEX to EMO is significant, reaching its maximum influence in the second phase and then gradually falling. In the VAR model (EMO D(INDEX)), lag order analysis shows that only lag 0 is significant. A suitable VAR model cannot be established for analysis. The results of the Pearson correlation of Rate(INDEX) and the sentiment of retail investors is 0.5159316953442752, and the p-value is 7.37838628287929e-10. The results show that the sentiment of retail investors and the change rate of the Shanghai Securities Composite Index are strongly correlated. Analysis of the TLCC shows that there is a lag in turn in the mutual influence relationship between the sentiment of retail investors and the change rate of the Shanghai Securities Composite Index. In the VAR model (sentiment & volume), the pairwise Granger causality test provides a p-value of 0.0121, which is significant. The null hypothesis (EMO does not Granger cause VOLUME) can be rejected. There is no causal relationship between the sentiment of retail investors and the Shanghai Securities Composite Index, and there is a causal relationship between retail investor sentiment and the total trading volume of the Shanghai Stock Exchange. There is a strong correlation and a mutual lag influence between the sentiment of retail investors and the rate of change of the Shanghai Securities Composite Index.

## *4.2Future research*

First, the crawler program needs to be further optimized since the existing program has the problem of insufficient stability. This problem becomes more prominent when the

target website itself is not stable enough. The stability and adaptability of the crawler program need to be further improved.

Second, with the advancement of big data technology, data sources should be more extensive in the future. In the future, the real-time and comprehensive advantages of big data technology should be incorporated.

Third, reinforcement learning algorithms in artificial intelligence, support vector machines and clustering calculations provide more diverse modeling options for affective computing algorithms. These methods have also achieved good results in some papers and provide more choices for affective computing in the future.

## REFERENCES

Ben-Rephael, A., Kandel, S. and Wohl, A., 2012. Measuring investor sentiment with mutual fund flows. Journal of financial Economics, 104(2), pp.363-382

Carpenter, J., F. Lu, and R. Whitelaw, 2015, The real value of chinese stock market, Working Paper (Stern School of Business, New York University).

Cho, V. and Wüthrich, B., 1999, April. Combining forecasts from multiple textual data sources. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 174-179). Springer, Berlin, Heidelberg.

Changsheng, H. and Yongfeng, W., 2012. Investor sentiment and assets valuation. Systems Engineering Procedia, 3, pp.166-171.

Chong, T.T.L., Liu, X. and Zhu, C., 2017. What explains herd behavior in the Chinese stock market?. Journal of Behavioral Finance, 18(4), pp.448-456.

Chen, C.C., Huang, H.H. and Chen, H.H., 2018. NTUSD-Fin: a market sentiment dictionary for financial social media data applications. In Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018).

Derakhshan, A. and Beigy, H., 2019. Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, pp.569-578.

Das, S.R. and Chen, M.Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management science, 53(9), pp.1375-1388.

Fernald, J. and Rogers, J.H., 2002. Puzzles in the Chinese stock market. Review of Economics and Statistics, 84(3), pp.416-432.

Fang, M., Yang, S. and Lei, Y., 2019. Residual contagion in emerging markets:'herd' and 'alarm'effects in informatization. Electronic Commerce Research, pp.1-21.

Fama, E.F., Random Walks in Stock Market Prices' (1965). Fin. An. J., 21, p.55.

Gazoni, E. and Clark, C., 2018. openpyxl-A Python library to read/write Excel 2010 xlsx/xlsm files. Retrieved from. (accessed on 20 April 2016) http://openpyxl. readthedocs. org/en/default.

Guo, K., Sun, Y. and Qian, X., 2017. Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. Physica A: Statistical Mechanics and its Applications, 469, pp.390-396.

Huo, X. and Qiu, Z., 2020. How does China's stock market react to the announcement of the COVID-19 pandemic lockdown? Economic and Political Studies, pp.1-26.

Hung, J.C., 2009. Deregulation and liberalization of the Chinese stock market and the improvement of market efficiency. The Quarterly Review of Economics and Finance, 49(3), pp.843-857.

Hunter, J. and Dale, D., 2007. The Matplotlib User's Guide. Matplotlib 0.90. 0 user's guide.

Jaybhay, K.M., Argiddi, R.V. and Apte, S.S., 2012. Stock market prediction model by combining numeric and news textual mining. International Journal of Computer Applications, 57(19).

Junyi, S., 2015. Jieba Python Library.

Kong, D. and Wang, M., 2014. The manipulator's poker: order-based manipulation in the Chinese stock market. Emerging Markets Finance and Trade, 50(2), pp.73-98.

Misina, M., 2003. What does the risk-appetite index measure?.

Mei, J., J. A. Scheinkman, and W. Xiong, 2009, Speculative trading and stock prices: evidence from Chinese A B share premia, Annals of Economics and Finance 10(2), 225–255.

McKinney, W., 2011. Pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14(9).

Ni, Z.X., Wang, D.Z. and Xue, W.J., 2015. Investor sentiment and its nonlinear effect on stock returns—New evidence from the Chinese stock market based on panel quantile regression model. Economic Modelling, 50, pp.266-274.

Oliphant, T.E., 2006. A guide to NumPy (Vol. 1, p. 85). USA: Trelgol Publishing.

Oliveira, N., Cortez, P. and Areal, N., 2013, September. On the predictability of stock market behavior using stock twits sentiment and posting volume. In Portuguese conference on artificial intelligence (pp. 355-365). Springer, Berlin, Heidelberg.

Oliphant, T.E., 2004. SciPy Tutorial.

Peterson, R.L., 2007. Affect and financial decision-making: How neuroscience can inform market participants. The journal of behavioral finance, 8(2), pp.70-78.

Pearson-correlation-coefficient-statistical-guide., (2020).[ Pearson Product-Moment Correlation online] Available at: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php> [Accessed 13 November 2020].

Qin, D., 2011. Rise of VAR modelling approach. Journal of Economic Surveys, 25(1), pp.156-174.

Qiu, L. and Welch, I., 2004. Investor sentiment measures (No. w10794). National Bureau of Economic Research.

Ritter, J.R., 2003. Behavioral finance. Pacific-Basin finance journal, 11(4), pp.429-437.

Richardson, L., 2007. Beautiful soup documentation. Dosegljivo: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018].

Reitz, K., (2020). Requests Documentation Release 2.25.0 [online] Available at: < https://buildmedia.readthedocs.org/media/pdf/requests/latest/requests.pdf> [Accessed 13 November 2020].

Stake overflow.com. (2020) coursera python final project sentiment classifier. Available from: https://stackoverflow.com/questions/62914117/coursera-python-final-project-sentiment-classifier [accessed 20 Jun 2020]

Sheppard, K., 2012. Introduction to Python for econometrics, statistics and data analysis. Self-published, University of Oxford, version, 2.

Schumaker, R.P., Zhang, Y., Huang, C.N. and Chen, H., 2012. Evaluating sentiment in financial news articles. Decision Support Systems, 53(3), pp.458-464.

Sun, L., Najand, M. and Shen, J., 2016. Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance, 73, pp.147-164.

Tsai, I.C., 2017. Diffusion of optimistic and pessimistic investor sentiment: An empirical study of an emerging market. International Review of Economics & Finance, 47, pp.22-34.

Wang, X., Wang, X., Zhong, Z. and Yao, J., 2020. The impact of US–China trade war on Chinese firms: Evidence from stock market reactions. Applied Economics Letters, pp.1-5.

Wüthrich, B., Permunetilleke, D., Leung, S., Lam, W., Cho, V. and Zhang, J., 1998. Daily prediction of major stock indices from textual www data. Hkie transactions, 5(3), pp.151-156.

Xuan, X., 2020.A Superficial Study on the Causes and Characteristics of Co-Movement in Chinese and American Stock Markets during the Epidemic.

Yao, J., Ma, C. and He, W.P., 2014. Investor herding behaviour of Chinese stock market. International Review of Economics & Finance, 29, pp.12-29.

Zhang, D., Hu, M. and Ji, Q., 2020. Financial markets under the global pandemic of COVID-19. Finance Research Letters, p.101528.

Zhu, Y.L., Min, J., Zhou, Y.Q., Huang, X.J. and Wu, L.D., 2006. Semantic orientation computing based on HowNet. Journal of Chinese information processing, 20(1), pp.14-20.

# APPENDICES

## Table

TABLE 1– Example of positive words and negative words dictionary

| | |
|---|---|
| positive | |
| 涨 | rise |
| 赚 | gain profit |
| 猛 | rush to gain profit |
| 进 | buy |
| 冲 | rush to buy |
| 上 | price will go up |
| 长多 | buy |
| 护盘 | Government or institution investor protect market |
| 利多 | information in market which shows it is better to buy |
| 牛市 | bull market |
| ... | |
| | |
| negative | |
| 跌 | price of stock goes down |
| 亏 | loss money |
| 走 | sell |
| 怂 | scared to buy |
| 下 | price go down |
| 拨档 | When investors are long, the stock price falls, and the stock price is expected to continue to fall. |
| 崩 | stock market is out of control to go down. |
| 利空 | information which leads market go down. |
| 暴雷 | suddenly the bad news come out to public. |
| 熊市 | bear market |

Figure

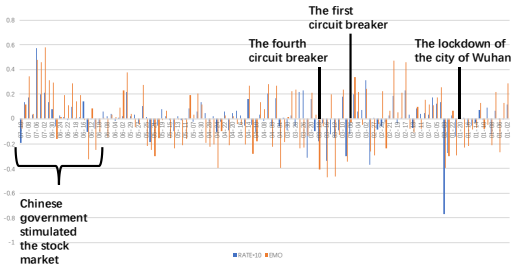



FIGURE 1 – Affective Computing Emotion Result of retail investors （EMO） and Shanghai Securities Composite Index Growth Rate (RATE)

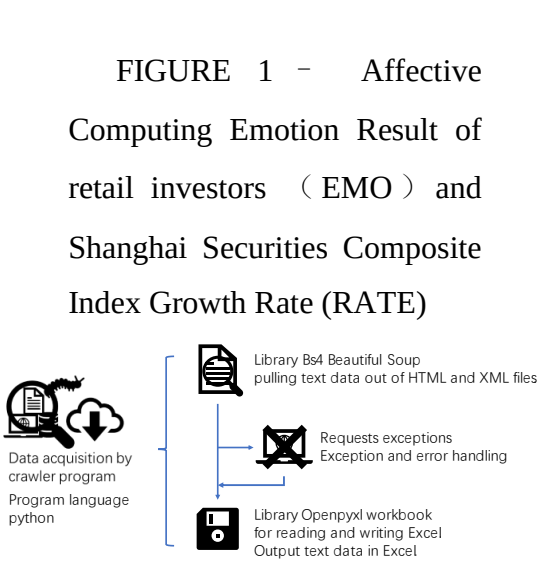



FIGURE 2 – Technical route.




FIGURE 3 – Structure of Crawler.

FIGURE 4 – Guba Shanghai Securities Composite Index Forum.

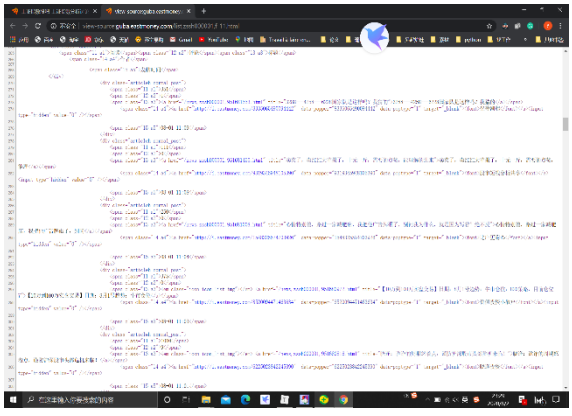


FIGURE 5 – Guba Shanghai Securities Composite Index forum source code.

```
from bs4 import BeautifulSoup
from openpyxl import Workbook
from requests.exceptions import RequestException
import re
import time
```

FIGURE 6 – Library called by the crawler.

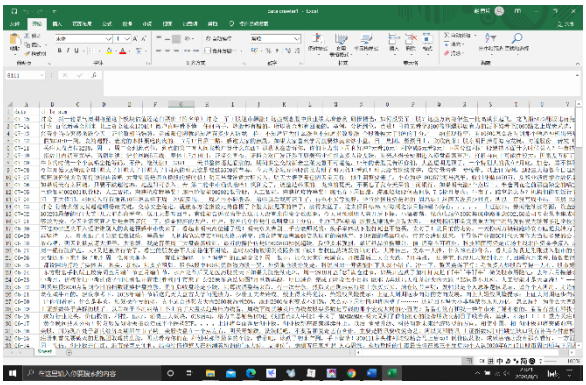


FIGURE 7 – Text data acquired by the crawler.

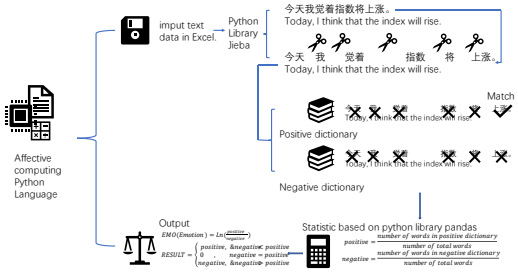



FIGURE 829 – The structure of affective computing.

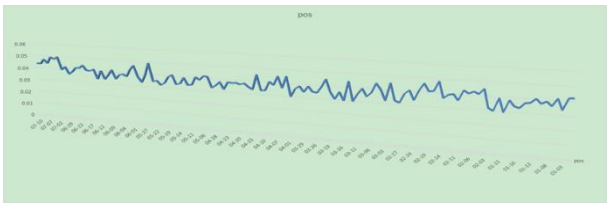


FIGURE 9 – Positive sentiment result.

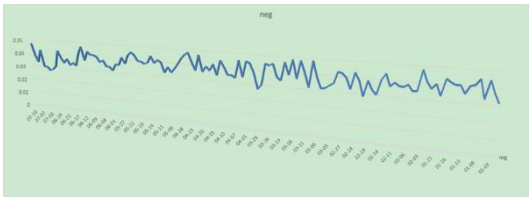


Figure 10 – Negative sentiment result.

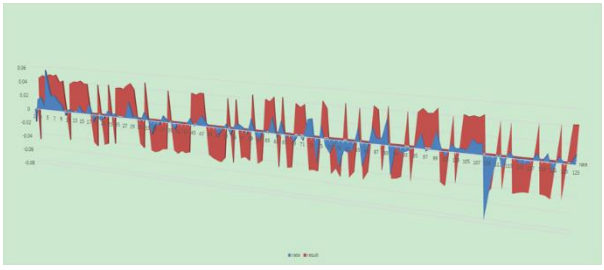


Figure 11 – Affective computing result （RESULT） and Shanghai Securities Composite Index change rate(RATE).

Roots of Characteristic Polynomial
Endogenous variables: EMO INDEX
Exogenous variables: C
Lag specification: 1 1
Date: 08/09/20 Time: 18:36

| Root | Modulus |
|---|---|
| 0.978915 | 0.978915 |
| 0.149239 | 0.149239 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

Figure 12 – EMO INDEX unit root test.

Roots of Characteristic Polynomial
Endogenous variables: EMO D(INDEX)
Exogenous variables: C
Lag specification: 1 1
Date: 08/15/20 Time: 14:37

| Root | Modulus |
|---|---|
| 0.163206 | 0.163206 |
| 0.025915 | 0.025915 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

Figure 13 – EMO D(INDEX) unit root test.

VAR Lag Order Selection Criteria
Endogenous variables: EMO INDEX
Exogenous variables: C
Date: 08/09/20 Time: 18:27
Sample: 1 125
Included observations: 117

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -730.0239 | NA | 932.1339 | 12.51323 | 12.56045 | 12.53240 |
| 1 | -563.8886 | 323.7509* | 58.32053* | 9.741686* | 9.883336* | 9.799194* |
| 2 | -561.9955 | 3.624381 | 60.46420 | 9.777701 | 10.01378 | 9.873548 |
| 3 | -559.3457 | 4.982654 | 61.88716 | 9.800781 | 10.13130 | 9.934966 |
| 4 | -558.1450 | 2.216584 | 64.94155 | 9.848633 | 10.27358 | 10.02116 |
| 5 | -554.8395 | 5.989428 | 65.75033 | 9.860505 | 10.37989 | 10.07137 |
| 6 | -552.0466 | 4.965142 | 67.17006 | 9.881139 | 10.49496 | 10.13034 |
| 7 | -550.1719 | 3.268708 | 69.72456 | 9.917469 | 10.62572 | 10.20501 |
| 8 | -547.2788 | 4.945490 | 71.14917 | 9.936390 | 10.73907 | 10.26227 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Figure 14 – EMO INDEX VAR Lag order selection criteria.

Vector Autoregression Estimates
Date: 08/09/20 Time: 18:31
Sample (adjusted): 2 125
Included observations: 124 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | EMO | INDEX |
|---|---|---|
| EMO(-1) | 0.151329 | 21.07640 |
| | (0.09438) | (18.4605) |
| | [ 1.60341] | [ 1.14170] |
| INDEX(-1) | 8.21E-05 | 0.976825 |
| | (0.00016) | (0.03074) |
| | [ 0.52220] | [ 31.7808] |
| C | -0.242087 | 70.19902 |
| | (0.46025) | (90.0257) |
| | [-0.52599] | [ 0.77977] |
| R-squared | 0.030699 | 0.905982 |
| Adj. R-squared | 0.014677 | 0.904428 |
| Sum sq. resids | 6.074675 | 232412.7 |
| S.E. equation | 0.224062 | 43.82655 |
| F-statistic | 1.916110 | 582.9958 |
| Log likelihood | 11.05311 | -643.1797 |
| Akaike AIC | -0.129889 | 10.42225 |
| Schwarz SC | -0.061656 | 10.49049 |
| Mean dependent | -0.001774 | 2928.811 |
| S.D. dependent | 0.225725 | 141.7661 |
| Determinant resid covariance (dof adj.) | | 52.78554 |
| Determinant resid covariance | | 50.26229 |
| Log likelihood | | -594.7666 |
| Akaike information criterion | | 9.689784 |
| Schwarz criterion | | 9.826249 |

Figure 15 – EMO INDEX VAR lag 1 model.

Pairwise Granger Causality Tests
Date: 08/09/20 Time: 18:23
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| INDEX does not Granger Cause EMO | 123 | 0.00388 | 0.9961 |
| EMO does not Granger Cause INDEX | | 2.11567 | 0.1251 |

Figure 16 – EMO INDEX Pairwise granger causality tests.

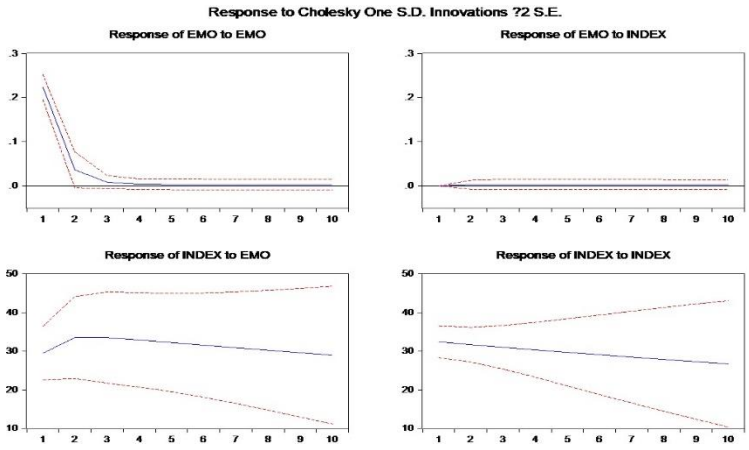



Figure 17 – EMO INDEX impulse analysis.

VAR Lag Order Selection Criteria
Endogenous variables: EMO D(INDEX)
Exogenous variables: C
Date: 08/10/20 Time: 13:21
Sample: 1 125
Included observations: 116

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -563.8150 | NA* | 59.12637* | 9.755430* | 9.802906* | 9.774703* |
| 1 | -561.4612 | 4.585717 | 60.83000 | 9.783814 | 9.926241 | 9.841631 |
| 2 | -559.6851 | 3.399174 | 63.21295 | 9.822157 | 10.05953 | 9.918519 |
| 3 | -558.0788 | 3.018714 | 65.88870 | 9.863427 | 10.19576 | 9.998334 |
| 4 | -554.6063 | 6.406160 | 66.51271 | 9.872522 | 10.29980 | 10.04597 |
| 5 | -552.1697 | 4.411091 | 68.36537 | 9.899477 | 10.42171 | 10.11147 |
| 6 | -550.8645 | 2.317839 | 71.67030 | 9.945940 | 10.56312 | 10.19648 |
| 7 | -549.1672 | 2.955591 | 74.64921 | 9.985642 | 10.69778 | 10.27473 |
| 8 | -546.3639 | 4.784970 | 76.30755 | 10.00627 | 10.81336 | 10.33391 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Figure 18 – EMO D（INDEX）VAR Lag order

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:31
Sample: 1 125
Lags: 1

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| D(INDEX) does not Granger Cause EMO | 123 | 0.00174 | 0.9668 |
| EMO does not Granger Cause D(INDEX) | | 0.30471 | 0.5820 |

Figure 19 – EMO D（INDEX）lag 1 Pairwise granger causality tests.

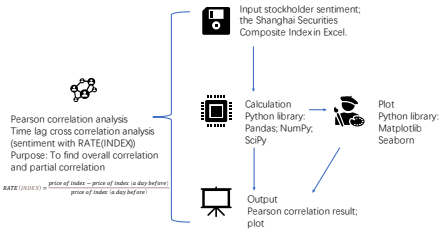



Figure 20 –The structure of Pearson correlation and Time lag cross correlation analysis

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:29
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| D(INDEX) does not Granger Cause EMO | 122 | 0.22761 | 0.7968 |
| EMO does not Granger Cause D(INDEX) | | 0.95721 | 0.3869 |

Figure 21 – EMO D（INDEX）lag 2 Pairwise granger causality tests.

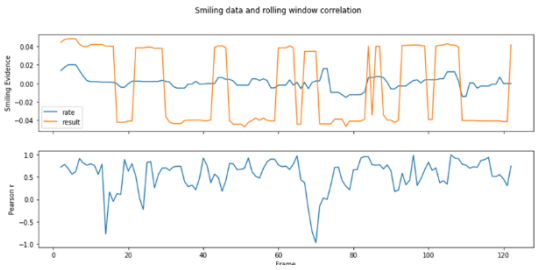



Figure 22 – Pearson analysis.

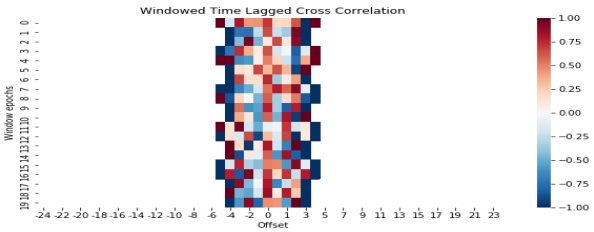



Figure 23 – TLCC analysis.

Roots of Characteristic Polynomial
Endogenous variables: EMO VOLUME
Exogenous variables: C
Lag specification: 1 2
Date: 08/10/20 Time: 13:45

| Root | Modulus |
|---|---|
| 0.975540 | 0.975540 |
| -0.480933 | 0.480933 |
| 0.348395 | 0.348395 |
| -0.173426 | 0.173426 |

No root lies outside the unit circle.
VAR satisfies the stability condition.

Figure 24 – EMO VOLUME unit root test.

VAR Lag Order Selection Criteria
Endogenous variables: EMO VOLUME
Exogenous variables: C
Date: 08/10/20 Time: 13:46
Sample: 1 125
Included observations: 117

| Lag | LogL | LR | FPE | AIC | SC | HQ |
|---|---|---|---|---|---|---|
| 0 | -6.330924 | NA | 0.003953 | 0.142409 | 0.189626 | 0.161578 |
| 1 | 99.54441 | 206.3212 | 0.000693 | -1.599050 | -1.457400 | -1.541542 |
| 2 | 110.2646 | 20.52413* | 0.000618* | -1.713925* | -1.477842* | -1.618078* |
| 3 | 114.0927 | 7.198049 | 0.000620 | -1.710986 | -1.380469 | -1.576800 |
| 4 | 115.8577 | 3.258508 | 0.000644 | -1.672781 | -1.247831 | -1.500257 |
| 5 | 115.8984 | 0.073844 | 0.000689 | -1.605101 | -1.085718 | -1.394238 |
| 6 | 119.7197 | 6.793370 | 0.000692 | -1.602046 | -0.988230 | -1.352844 |
| 7 | 124.6673 | 8.626595 | 0.000682 | -1.618245 | -0.909995 | -1.330704 |
| 8 | 129.0757 | 7.535727 | 0.000678 | -1.625226 | -0.822543 | -1.299346 |

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Figure 25 – EMO VOLUME var lag order selection criteria.

Vector Autoregression Estimates
Date: 08/10/20 Time: 13:42
Sample (adjusted): 3 125
Included observations: 123 after adjustments
Standard errors in ( ) & t-statistics in [ ]

| | EMO | VOLUME |
|---|---|---|
| EMO(-1) | 0.081984 | 0.153052 |
| | (0.09790) | (0.05100) |
| | [ 0.83742] | [ 3.00078] |
| EMO(-2) | 0.088792 | -0.026536 |
| | (0.09640) | (0.05022) |
| | [ 0.92104] | [-0.52834] |
| VOLUME(-1) | 0.245141 | 0.587592 |
| | (0.17562) | (0.09149) |
| | [ 1.39586] | [ 6.42221] |
| VOLUME(-2) | -0.152757 | 0.364912 |
| | (0.17460) | (0.09096) |
| | [-0.87489] | [ 4.01163] |
| C | -2.446018 | 1.265450 |
| | (2.29063) | (1.19336) |
| | [-1.06783] | [ 1.06041] |
| R-squared | 0.063711 | 0.840198 |
| Adj. R-squared | 0.031972 | 0.834781 |
| Sum sq. resids | 5.851959 | 1.588312 |
| S.E. equation | 0.222695 | 0.116018 |
| F-statistic | 2.007353 | 155.1032 |
| Log likelihood | 12.76315 | 92.96559 |
| Akaike AIC | -0.126230 | -1.430335 |
| Schwarz SC | -0.011914 | -1.316018 |
| Mean dependent | -0.002827 | 26.44295 |
| S.D. dependent | 0.226342 | 0.285428 |
| Determinant resid covariance (dof adj.) | | 0.000555 |
| Determinant resid covariance | | 0.000510 |
| Log likelihood | | 117.1328 |
| Akaike information criterion | | -1.741996 |
| Schwarz criterion | | -1.513364 |

Figure 26 – EMO VOLUME VAR model.

Pairwise Granger Causality Tests
Date: 08/10/20 Time: 13:39
Sample: 1 125
Lags: 2

| Null Hypothesis: | Obs | F-Statistic | Prob. |
|---|---|---|---|
| VOLUME does not Granger Cause EMO | 123 | 1.24079 | 0.2929 |
| EMO does not Granger Cause VOLUME | | 4.58500 | 0.0121 |

Figure 27 – EMO VOLUME VAR model pairwise granger causality tests.

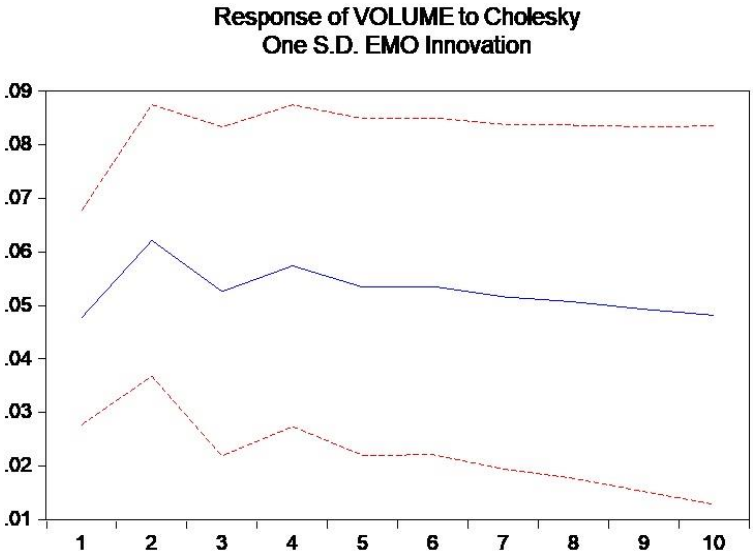



Figure 28 – EMO VOLUME impulse analysis