Sentiment Analysis and Stock Price Movement: Shanghai Stock Exchange - A Case Study

Zihan Huang

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Intelligent Systems)

Supervisor: Khurshid Ahmad

August 2019

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Zihan Huang

August 9, 2019

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Zihan Huang

August 9, 2019

Acknowledgments

I would like to thank a number of people who have helped make this dissertation possible.

I am deeply thankful to my supervisor, Professor Khurshid Ahmad, who has guided me through the formative process of this dissertation and offered me his continued support, expertise, valuable feedback and encouragement.

I would also like to thank Trinity College Dublin, the School of Computer Science and Statistics, for the recognition and opportunity to learn from the very best in the field of Computer Science.

Most of all, I am immensely grateful for my family, who without their unconditional love and support I would most certainly not be here today.

Zihan Huang

University of Dublin, Trinity College
August 2019

Sentiment Analysis and Stock Price Movement: Shanghai Stock Exchange - A Case Study

Zihan Huang, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Khurshid Ahmad

Behavioural finance suggests and proves that human behaviours play a vital part in market, which could be one possible explanation to market acting abnormal. Meanwhile, people depend on information to make decision which leads to possible price changes, which means the information source plays a vital part in determining peoples attitude towards the topic.

The main objective of this dissertation is to combine methods from the area of textual analysis and time series modelling to process and analyse the sentiment from news source text to determine a relationship with movements in assets traded on financial markets. As for sentiment, this dissertation take negative sentiment and oral sentiment, the sentiment proxy for each is calculated using textual analytics techniques and using an extra dictionary built by this dissertation.

I have developed a method for analysing the impact of news on the price of a stock or the value of an index of stock prices. The term news here refers spastically to the frequency of mentioning name entities in a market, and potential sentiment expressed about the entity. My method contains crashing a time series of price and the fusion with an equivalent time series of sentiment in the news and/or time series comprising the frequency by which name entity is cited in the news. For fusion I have used vector autoregression (VAR) which helps to examine the impact of news on the price

movement, e.g., a share or index. Shortly, I have comprised approximate 18000 news text from a large news agency (Xinhua) along with Chinese stock price from Shanghai Stock Exchange (SSE). Using sentiment proxies, sentiment series were generated after text analysing and aggregated with return series processed from price series, and I have some evidences showing that a correlation between sentiment and price, including negative and oral sentiment of small and median enterprises in Shanghai Stock Exchange.

Summary

The main objective of this dissertation is to determine and investigate the relationship between news and price movements in assets traded on financial markets.

I have developed a method for analysing the impact of news on the price of a stock or the value of an index of stock prices. The term news here refers spastically to the frequency of mentioning name entities in a market, and potential sentiment expressed about the entity. My method contains crashing a time series of price and the fusion with an equivalent time series of sentiment in the news and/or time series comprising the frequency by which name entity is cited in the news. For fusion I have used vector autoregression (VAR) which helps to examine the impact of news on the price movement, e.g., a share or index. Shortly, I have comprised approximate 18000 news text from a large news agency (Xinhua) along with Chinese stock price from Shanghai Stock Exchange (SSE). Using sentiment proxies, sentiment series were generated after text analysing and aggregated with return series processed from price series, and I have some evidences showing that a correlation between sentiment and price, including negative and oral sentiment of small and median enterprises in Shanghai Stock Exchange.

The work can be separated into three stages, collect and process both textual data from news source and financial time series data, aggregate financial time series with time series of sentiment and build statistical models. Besides negative sentiment, this dissertation also will be investigating the impact from oral sentiment of Chinese

company with different size in news on price movement.

To achieve the research goals economics-oriented financial news pieces were collected through a news vendor called LexisNexis over a three-years period. The collected texts were aggregated to a sentiment time series calculated and generated based on the text using the Rocksteady affect analysis system. Meanwhile, a financial time series was also retrieved and processed with logarithmic calculation and output with a return time series for the same period of time. Later on, two time series were then aggregated and aligned by date and time and examined using four different vector autoregressive (VAR) models, which individually examine the independent effect of different exogenous variables such as sentiment and different sized companys density on financial returns.

An analysis of approximately 18000 articles about Chinese financial market, the focus of the case study, collected for the period from January 2015 to May 2019. The study also introduced three variants representing key events happened in the period of time this dissertation looked at, by doing so the extra sentiment of uncertainty is balanced out and prevent the modelling from over fitting.

The study carried out found that negative sentiment plays a small role in explaining returns of an asset, however, the significance it showed on return is consistent. It shows that the explanatory power of sentiment has a delay effect as in different lagged terms show different in terms of level of significance on return. The oral sentiment of company shows an interesting impact too, companies are categorised into three: large, median and small. From the modelling, large company showed no impact on return, but small and median company showed a relatively small, delayed but consistent statistical significance on returns.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The observation on market price movement shows that there are time periods where price does not act according to supply and demand: in stock market, when price goes up showing a profit in selling, shareholders see the profit and sell more which leads to price going down because the supply is more than demand, and vice versa. There are times where price keeps going up; shareholders keep buying knowing the price is high, and where price keeps dropping; the shareholders keep selling with a dropping price. For example, the Shanghai Exchange composite index price movement shows ups and downs (see Figure 1.1).

The abnormality leads to investigation what could be the reason behind it, how can it be explained. One possible potential explanation lies in human behaviour; the use of sentiment and emotions when making decisions on selling and buying.

Figure 1.1: Four years of Shanghai Stock Exchange Composite index adjusted close price during the four-year period (2015-2019): the market index was at its peak (at 5166.4 on 12th June, 2015) and has seen major downturns few months later (at 3507.2 on 8th July, 2015, at 2965 on 25th August, 2015 and at 2655.7 on 28th January, 2016).

The economists have developed the so called efficient market hypothesis that the prices of financial assets are a reflection of all available information, and that in general, competition will cause the full effects of new information on intrinsic values to be reflected instantaneously in actual prices [1].

With the appearance of newspapers and social media in other form, the way information evolved from in person communication: for example, word of mouth, to read from published news piece or articles. As the purpose of advertisement on television is to appeal viewers with their products, it is natural to assume that opinions can be influenced by news in various media format: televised, digitalised or in print media.

How newspapers that may have information that has impact on investors decisions is one of the interests behind this paper. The flow of information and uncertainty in

markets influences price discovery, price changes, volatility, the behaviour of market participants, and market stability.

I will be exploring the use of sentiment proxy which is the number of words used usually denote negative feelings or sentiments. Another sentiment proxy is news flow that is the number of news items per day based on the expression of news items goes up only in time of crisis and settle down to average figure afterwards. Further, firm-level proxy is also used in combination with sentiment proxy when doing the statistical modelling, which is the number of company name mentioned and it categorised into large, median and small company.

The main objective of this dissertation is to determine and investigate the relationship between news and price movements in assets traded on financial markets. Methods developed including process data collection, text analytics on analysing the sentiment from news source text and statistical modelling on using combined methods from the area of textual analysis and time series modelling.

1.2 Key Conclusion

The contributions of this thesis lay in the market that be applied the method on which are used to construct the sentiment proxy and evaluate its effectiveness as an explanatory variable in a statistical model. Previous authors have studied equity markets and text-based sentiment; a number of studies have looked at the firm-level proxy in related to the market when evaluating sentiment. Comparably few studies have focused on Chinese markets and sentiment from Chinese newspapers. Results presented in this study verify and evaluate the method and find a similar significant impact of news sentiment on the returns of the Shanghai Stock Exchange index.

This study examined the correlation between negative sentiment, oral sentiment of companies and price movement which is indicated by return values. Findings on Chinese market showed that sentiment proxy always carries a significance on return which is relatively small but consistent. Additionally, the study found that oral sentiment of small company and median company show significance on return as well, which are smaller than it of sentiment proxy but consistent. However, large company density shows no significance on return. All observed impacts are delayed effect which is in consistency with previous study.

1.3 Thesis Structure

The approach includes three stages of work: the retrieval and processing of financial data and corpus of news data, textual sentiment analysis on corpus and statistical modelling of financial and sentiment series.

The next chapter introduces sentiment analysis, including definitions of sentiment analysis on different subjects, sentiment analysis case studies on index level and firm level and investigation in natural language processing (NLP) solution for sentiment analysis.

The following chapter gives details of the methods of this work, the outline of each phase work. Following the methods, case study is explained in details, and results are shown and discussed.

Based on the case study conducted, conclusions are made in the final chapter and future work worth pursuing is proposed.

1.4 Conclusion

This study takes a special interest in investigating how sentiment is related to market price movements as in how human behaviours can contribute and result in changes in price. The special interest of this dissertation is to take focus at Chinese market with sentiment of negativity and oral sentiment taking Chinese company as the target with proxy representing the frequency of name entity mentioned in the text source with regards to Chinese market, and a case study was done on Shanghai Stock Exchange. Results showing that the impacts can be observed in Shanghai Stock Exchange index and in consistency with previous study, additionally, interesting observation on the impacts shown from oral sentiment proxies on price (See Chapter 3,4).

Chapter 2

Motivation and Literature Review

2.1 Introduction

Previous discussions introduced the object of this work is to exam the possible explanation of market misbehave with human sentiment, and how to extract sentiment from textual data is one of the focuses in this study. Here I will introduce the definition of sentiment analysis in regards to different domains that this work covers, also focus on how sentiment is related to market behaviour based on previous research in domain of behavioural finance. As sentiment is extracted from textual data, techniques available to do natural language processing are discussed.

This chapter includes studies that were published on the subject of sentiment analysis in the domain of finance and text analysis in particular, during the time from 1990 to present. Studies on sentiment and price are reviewed and discussed on index level and firm level. The solutions for natural language processing in relation to sentiment analysis are discussed on lexical level approach and machine learning approach, additionally the solution adopted in this work is discussed.

This chapter will begin with presenting different definitions of sentiments analysis regarding various domains (see Section 2.2). Following is detailed discussion on chosen researches which applied sentiment analysis on different levels (see Section 2.3). Last but not the least, existing natural language processing (NLP) solutions for sentiment analysis will be discussed.

2.2 Some Definitions

2.2.1 Sentiment Analysis as a NLP problem

Sentiment refers to what one feels with regards to something; mental attitude (for example, of approval or disapproval); an opinion or point of view as to what is right or agreeable (according to the Oxford English Dictionary). For sentiment, sentiment is psychological constructs and difficult to measure in archive analysis, how to identify and quantify one or more sentiments is the aim of sentiment analysis.

Sentiment analysis, or opinion mining is defined as algorithmic classification of users evaluations of a brand (positive, negative, or neutral) inpostsand comments (according to the Oxford English Dictionary). Sentiment analysis refers to the method that transfers unstructured textual contents to structured data[2, 3], including the use of natural language processing, text analysis and computational linguistics [4].

2.2.2 Sentiment Analysis in Behavioural Finance

In theory, the price of a certain item should be determined by its supply and demand. However, in the market, price sometimes changes out of the blue and usually is affected by multiple reasons, for example, human behaviours. Among the attempts to explain the movements of stock market price, sentiment was shown to be one of the factors that contributes to the price changing.

Behavioural finance studies the impact on the decision-making area and the effect on market price, return and the risk of market from psychological or other human factors; it has demonstrated that investment decision of investors is more likely driven by their emotions [5].

Research in the domain of psychology proved that emotions with additional information has impact in human making decision, and the impact is not subtle[6].

Previous works had shown that negative sentiments: such as anxiety, worry and fear mood, have indications to down-ticks in stock price in the short future [6]. Researchers had found that sentiments can be used to predict price movements [4, 7]: there is correlation shown between the collective mood states from large-scale twitter feeds and the value of the Dow Jones Industrial Average (DJIA) over time, and emotions can be used to predict NASDAQ and S&P500 index.

Sentiment regarding market price is categorised into, *market sentiment* and *consumer sentiment*. Market sentiment reveals the overall attitude of investors toward a particular market, which is also called *investor sentiment*. *Consumer sentiment*, or *consumer confidence*, is determined by consumer opinion: for example, ones feelings toward ones own current financial health, the health of the economy or market (see Figure 2.1 as example). Note that *consumer confidence* is gathered at a much lower frequency.

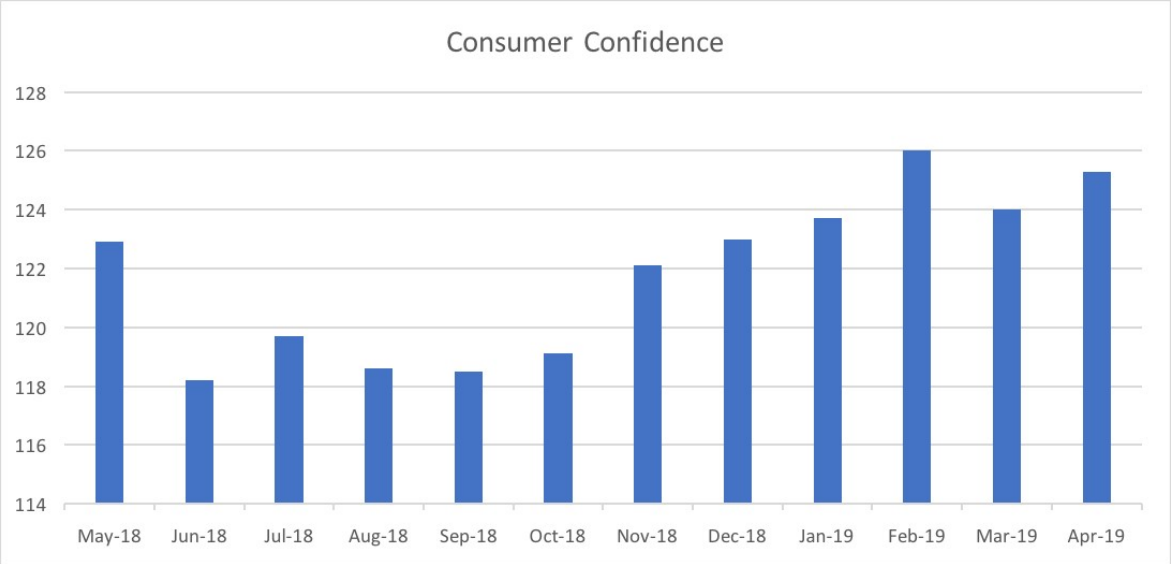

Figure 2.1: Chinese consumer confidence. The Chinese consumer confidence index is calculated from the results of a survey taken by 70 individuals over 15 years old from twenty cities over the country, covering the consumer expectation and consumer satisfaction index. The confidence index measures and concludes the degree of satisfaction about the current economic situation and expectation on the future economic trends, from 0 to 200 to be extreme pessimism to extreme optimism.

2.2.3 Sentiment Analysis in Changing Market

Stock market price movement is of interest among researchers, however, previous studies are mostly looking into the US and EU markets, for example, using sentiment analysis on news, stories and opinion extracted from Twitter. There has not been much research done aiming at Chinese markets or Chinese social media, such number

of papers is limited, and the number of researches aimed at applying sentiment analysis to inspect Chinese stock market is even lower despite the fact that China has already become the second largest economies in the world.

Chinese stock market is a merging market and it differs from mature markets such as the US market in few aspects, the most noticeable one is the structure of market participants. China Securities Depository and Clearing Statistical Yearbook in 2013 noted that 99.5% of the total investor accounts are individual investors accounts which takes up more than 80% of the market. However, in the US market, over 50% of trading are made by institutional investors [8]. It shows that Chinese stock markets are dominated by individual investors who are less likely to have enough knowledge and experience in analysis the market. It is riskier to trade in these stock exchanges since the prices are more easily controlled by special events and public moods, meaning that individual investors are more likely to decide on buying or selling stock based on what they have heard or seen from news or other social medias[9].

2.3 Sentiment Analysis Case Study

In the study of financial market, some scholars distinguished between the noise traders and rational traders. I am going to briefly review how the notion of sentiment analysis enters modern literature of finance and computing. I will begin with one or more key authors from 10 years before the last century (1990-2000). Then I will look at the stock market works carried out on share price, index where the investors behaviour is related to items in the press or in social media (2000-2010). Later on, people take sentiment analysis based on sentiment expressed on social media data stream and with calculation. All the above studies relate to stock market, or more precisely, a stock market index. Recently, some studies have been done on firm-level, that is a single firm comprise the price sentiment are picked on from news articles (2016 to present).

Table 2.1: Researches on the topic of financial market and sentiment with different focus from 1990 to present. (*: the term social mood is supposed to comprise feelings and emotions: terms which are related to sentiment)

Period	Author(s)	Financial Data	Sentiment Proxy
1990-2000	De Long J. B. *et al.*	Market	(theoretical paper)
2000-2010	Baker M. *et al.*	IPO's price changing	Price Movement
	Paul C. Tetlock	Stock market index	Wall Street Journal
	Kling G. *et al.*	Share index	Investor sentiment
	J.R. Nofsinger	Merging Market	Social mood*
2011-2015	Bollen *et al.*	DJIA	Twitter
	Zhang X. *et al.*	Stock market index	Twitter
	Bhardwaj, A. *et al.*	Stock market index	Sensex, Nifty
2016-*present*	Ahmad, K. *et al.*	Firm-level index	News Articles

2.3.1 Noise Traders Risk

In the study of financial market, some scholars distinguish between the noise traders and rational traders. Noise traders behaviours pose a risk to market which cannot be understood in the framework of economics: it does not deal with irrational behaviours[6]. Scholars argue that the investor can be categorised into *nave* and *expert*, which represent unpredictable, unsophisticated and *noise traders* and the contrarian, *rational traders*; and some financial market anomalies can be explained by the idea of noise trader risk. Their essential assumption is that the opinions of noise traders are unpredictable and the risk that their misperceptions would become more extreme the next day requires the rational traders in the market to bear. Scholar build model to further show and prove their hypothesis that the behaviour of expert traders is a response to noise traders rather than on fundamentals. Their analysis proved the point that analysing the irrational behaviour does not always require specifying its content, which can be done simply looking at the effect of unpredictability of irrational behaviour on the opportunities of rational investors.

2.3.2 Market Index and Social Mood

On observing few downticks of firm stock price after sold as IPO, researchers challenged on the classical finance theory: researchers argue that investor sentiment has no impact on stock prices or returns, and take special interest in to Initial public offering (IPO).

Finding shows that the cross-section of future stock returns is conditional on beginningofperiod proxies for sentiment [10]. Stocks (such as younger stocks, small stocks, unprofitable stocks, high volatility stocks, etc.) that are attractive to optimistic investors and unattracted to arbitrageurs tend to earn relatively low subsequent returns.

Conditionally on low sentiment, however, these cross-sectional patterns attenuate or completely reverse. As the results cannot be fully explained by classical theory since the results reflect the compensation for systematic risks, from then on, more researches took their stand on the inconsistency.

Later on, as investors sentiment be found having correlation with price movement, further examination showed that institutional investors sentiment has no long-run relation with stock prices but short-term dynamic matters [8]. Research shows that market returns influence investors sentiments, but not vice versa as in institutional investors sentiment does not predict future stock returnsbut stock returns influence sentiment in the sense of a positive-feedback process.

Since behavioural economics argues that emotions can profoundly affect individual behaviour and decision-making, question arose in if it is the same apply to large societies: societies experience mood states that affect their collective decision making? Researchers take Twitter feed as public mood source and investigate the possible correlation between it and share prices. Results indicate that the accuracy of prediction of share price (for example, DJIA) can be significantly improved by the inclusion of specific public mood dimensions but not others [7].

2.3.3 Market Index and Sentiment from Social Media

Entering the era of social media, researchers take a special interest into sentiment extracted from social media: for example, research find that emotional tweet percentage significantly negatively correlated with Dow Jones, NASDAQ and S&P 500, but displayed significant positive correlation to VIX [4]. It therefore seems that just checking on twitter for emotional outbursts of any kind gives a predictor of how the stock market

will be doing the next day[7].

Using Wall Street Journal reports as source of sentiment, Tetlock [11] implemented the sentiment analysis on the text from opinion column on Wall Street Journal and quantifies the emotions into optimism and pessimism. His observation from the results was that trading volume tends to increase after pessimism reports and high pessimism scored reports tend to be followed by a down trend and a reversion of market prices.

In addition, Tetlock showed in his work that the negative affective component of news has a longer lasting impact on volatility of stocks. In another paper[12], Tetlock *et al.* use Harvard IV-4 psychological dictionary for exploiting emotions into two dimensions, positive and negative. They used the dictionary to analyse the fraction of negative words in Dow Jones News Service and Wall Street Journal stories about S&P 500 firms from 1980 through 2004. Their conclusion is that the content from media captures information of different aspects of firm, which are hard to quantify in other ways, and investors are able to be exposed and affected by that to make decisions on purchasing.

2.3.4 Firm-Level Index and sentiment from newspaper

In 21 centuries, Ahmad *et al.*[13] investigated the relation between media-expressed firm-specific negative tone and firm-level returns. Their corpus includes more than 5.5 million articles from newspapers, industry and trade magazines, financial blogs and so on. As for firms, they took 20 large non-financial firms in the US and took daily data over ten-year period of 2001-2010 for each. By estimating a series of separate rolling window vector autoregressive (VAR) models for each firm, they show how media-expressed negative tone impacts firm-level returns episodically in ways that vary across firms and over time.

Their finding shows that a rise in firm-specific negative media tone leads to a significant next-day firm-level returns but followed by a reversed movement right after that, and firm-specific negative tone impacts on firm-level returns occasionally. The observed lagged return effects in their time-varying Vector Auto regression models are explained that it may take time for the market to process the information and to reflect. Their findings are in general consistent with efficiently functioning markets in which the media assists with the processing of complex information.

2.4 NLP solutions for Sentiment Analysis

For machines to specify and quantify the subject, for example, opinion and sentiment, within textual content such as a piece of news, techniques including natural language processing (NLP) are employed [14], the solutions can be broadly categorised into machine learning approach and lexicon-based approach.

2.4.1 Lexicon-based Approach

Based on the idea that the level or strength of a sentiment in a piece of text can be build up by summing up the sentiment orientation of different used words or particular phrases in the text, lexicon-based approach is developed[15]. There are two more specific approaches under lexicon-based category depending on how to quantify the sentiment orientation: corpus-based and dictionary-based.

Depending on the co-occurrence patterns of words in the text, the corpus-based approach determines whether a sentiment is present on the corpus itself without any modification or processing to break the context.

With a list of opinion word/phrases called seed list, the sentiment of the text is determined by what other opinion-oriented words/phrases having similar context to the seed list has found in the text[16].

However, the dictionary-based approach breaks up the text by words or phrases, and uses a predefined dictionary where each word is assigned with a specific polarity strength according to linguistics and psychology [16]. How it works is that each piece of text is broken down into single word as token (tokenisation) and the frequency of each word is counted, then scores for different sentiment category are computed by the frequency or percentage of tokens that are matched with the affect categories in the dictionary employed [17]. While choosing the dictionary to use, the common approach is adopting a developed lexicon (for example, the General Inquirer dictionary, known as a comprehensive English language dictionary, is generated merging the Harvard IV-4 and Laswell dictionary).

2.4.2 Machine Learning Approach

As part of the problem solving, classification is a key technique to process the textual data. Binary classification models are usually used when processing textual data, for instance, positive or negative, happy or sad. As technology developed, modern approaches in the field of machine learning are become popular such as Nave Bayes, Support Vector Machine (SVM) and so on.

SVM is one of the supervised machine learning algorithm family. Basically, it learns the features of sets of labelled data and classified data to different sets. As a result, it output an optimal hyper-plane as boundary between different labelled data.

Nave Bayes is a classification algorithm with statistical approaches which has the assumption made that each features of the instance data is independent. How it works is that to assign the label to instance data, as in determine which class that the instance belongs.

Different from lexicon-based approach, machine learning techniques don't depend on a database as such in general. However, the drawbacks lie in that fact there are insufficient authenticated labelled data resources for training models at the first place, and if it does, the large amount of data for training could take too much time.

2.4.3 Rocksteady

This dissertation has used Rocksteady as sentiment analysis tool which employs the dictionary-based approach under the former category discussed.

Rocksteady is an affect analysis system developed by Ahmad *et al.* at Trinity College Dublin. Rocksteady deploys a dictionary-based approach and comes with a general purpose dictionary, and it also allows user to add a custom dictionary adding on to the base one which provides more opportunity for researchers to look at a specific area for analysis purpose.

2.5 Motivation

As price should be changing to supply and demand in markets: when price goes up, shareholders see the profit and will sell and when there are more selling than buying, price will go down; when price is low, shareholders will more likely to buy than selling

which leads to price rising. However, from observation on Shanghai Stock Exchange composite index, there are times when price kept going down, meaning shareholders decide to sell even though the price is low already, same observation found in other markets as well.

From external information, shareholders analyse the current situation of market and make judgments on possible trend of the market which leads to their final decisions on buying or selling: how they feel about the markets decided if they buy or sell; that is sentiment playing a part in price changing.

The motivation behind this work is to take sentiment analysis techniques into modelling financial returns. This work investigated into different sentiment analysis techniques and impact of sentiment on price: taking sentiment as a variate into multivariate time series analysis of stock price return.

This dissertation is to fill the gap of literature in the he field of finance and computing that focuses on Chinese market with Chinese formal media as sentiment source to investigate the impact on price movement. Chinese market is a merging market and it may reveal a difference from observations on EU or US market. Plus, comparing to social media, which were of researchers interest for the past decades, formal media are less likely to be proven as fraud. Also, this work took a special interest in investigate the impact of oral sentiment on price movement from news, taking the frequency of name entity mentioned in the news source. In this case, part of this dissertationẃork is to create a dictionary that categorised different size of companies and added to text analytics to calculate the proxies for orla sentiment of companies.

2.6 Conclusions

This chapter, has presented an overview of literature from a cross section of studies in sentiment analysis, NLP solutions for sentiment analysis and definition of sentiment analysis in different circumstances. Different NLP approaches are discussed and focused on the domain of finance, including case study covering sentiments analysis on index level and firm level. The methodology behind each NLP solution is discussed and the system this study use is introduced.

From the literature review, this study has a thorough understanding of existing techniques and different focus of previous researches and decided on choosing Chinese

market as the focus with official newspapers as sentiment source. System design, data acquisition and processing and statistical modelling will be introduced next (Chapter 3).

Chapter 3

Methods

3.1 Introduction

Here I will introduce the methods of how the study carried out. Basically, the process was divided into three parts: data acquisition and pre-processing, sentiment analysis on textual corpus and statistical modelling. The architecture (see Figure 3.1.) will give a clear view of how each part connects to each other, base on which I will explain more in details in the following subsections

3.2 Architecture

The study carried out in three stages, including financial data and textual data retrieval, sentiment analysis and statistical modelling.

First stage of the study is data collection, both financial data and formal media data (news) was collected from online with open source access, Yahoo Finance as financial data source and LexisNexis as social media data source. With raw time series data, processing needs to be done in order for the study to carry on, financial data is going through calculation and become return time series while textual data is being analysed using rocksteady and extracted with sentiment time series. Later on, two time series will be aggregated and ready for statistical modelling.

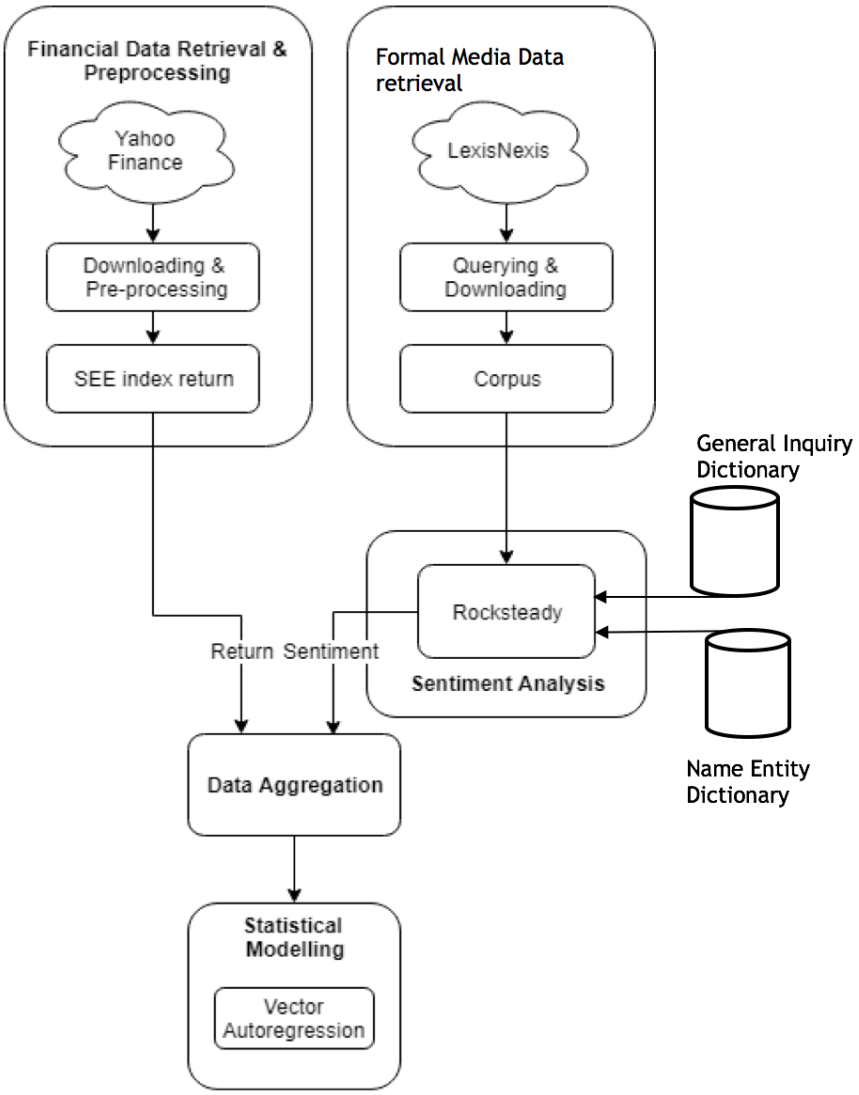

Figure 3.1: System architecture overview. The approach followed is, firstly collecting data from trusted sources granted with access permission. After text analysis on news series, sentiment time series is generated and aggregated with return time series. Statistical modelling and evaluation of the models are carried on in the end.

3.3 Data Acquisition and Pre-processing

In this section, methods used to collect data is introduced. With collected raw data, pre-processing done for later aggregation is discussed in details. The data needed for this study can be categorised into financial data as source of return data and textual corpus as source of sentiments which will be discussed separately in detail.

3.3.1 Financial Data

As discussed previously, the market that this study takes focus on is Chinese market. The stock market in mainland China has grew into one of the largest stock market globally, which includes Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE). This study takes focus Shanghai Stock Exchange in particular, to observe and investigate the asset price.

There are two types of stocks are traded in SSE: A shares and B shares (A shares are denominated in Chinese currency, and B shares in the contrary are settled in foreign currencies); Shanghai Stock Exchange composite index (SSE index) is a capitalization-weighted index and it tracks the daily price performance of all A-shares and B-shares listed and this study will be using it as the financial equity.

As for source of historical price data for SSE index, this work use Yahoo Finance. Yahoo! Finance is part of Yahoo! network which mainly provides financial news, data and commentary including stock quotes, press release and original content. It is a perfect source to obtain historical data for SEE index with only one limitation that it only has previous five-years data of SEE index.

In this study, SSE index data from January 2015 to April 2019 was collected, with approximate 250 of daily data per year: weekends and public holidays were excluded; the daily index data included the open price, close price, daily highs, lows and trading volume. Upon obtaining historical price data of SEE index, daily price is highly correlated and not optimal for investigating (see Table 3.1).

Table 3.1: Four years of Shanghai Stock Exchange Composite index correlation example; correlation with one to five day(s) of lags are listed in the table. Note that the correlation with one-day lag is 99.2% which means todays price is 99.2% correlated to previous days price, and the correlation decrease to 95.7% with five-days lag which means that todays price is still highly correlated to the price five days ago.

Lag(s)	1	2	3	4	5
Correlation	99.2%	98.3%	97.6%	96.7%	95.7%

According to financial texts one should focus on the change in prices rather than on the price itself [2]: typically, it is suggested that we take a ratio of the price which then discounts for currency fluctuation and other changes. Frequently, the logarithmic return value is used (see Equation 3.1):

$$r(t) = log(p(t)/p(t-1)) \tag{3.1}$$

The nature of logarithm can perfectly represent the changes in price: for example, of the day and of the day before, p(t) and p(t-1). If the price was the same on two successive days, then the ratio of two days returns will be one and the return will be zero (see Equation 3.2).

$$r(t) = log(p(t)/p(t-1)) = log(1) = 0 \tag{3.2}$$

If the price is higher than the previous day, then the ratio will be greater than one the return will be positive. If the ratio is less than one, then the log will be a negative number (see Table 3.2).

Table 3.2: Shanghai Stock Exchange composite index return samples. Returns shows negative values and positive values which indicate how the value change between days and previous days: for example, p(t) on 1/7/2015 is greater than it on 1/6/2015, so r(t) is positive, while p(t) on 1/8/2015 is less than it on 1/7/2015, return is negative.

t	1/5/2015	1/6/2015	1/7/2015	1/8/2015	1/9/2015
p(t)	3350.5	3351.4	3374.0	3293.5	3285.4
r(t)	-	0.00028	0.00669	-0.02415	-0.00245

The fact that return is positive means that the price goes up comparing to the day before; price going up shows a profit in selling while it shows a lose if the return is negative. When price goes up, shareholders see the profit and sell more; price changes in supply and demand so when there are more selling than buying, price goes down. The series appear have one uptick for every downtick that is the right price has been discovered.

Table 3.3: Four years of Shanghai Stock Exchange Composite index returns correlation example; correlation with one to five day(s) of lags are listed in the table. Note that the correlation with one-day lag is 5.25% which means todays price movement is 5.25% correlated to previous days price movement, and the correlation decrease to 1.07% with five-days lag. Recall the correlation of prices with one day lag is 99.2% and for five days of lag it is 95.75%, the correlation between return is much lower than it of price.

Lag(s)	**1**	**2**	**3**	**4**	**5**
Correlation	5.25%	-6.24%	2.43%	8.01%	1.07%

Consistent with financial text, price return shows price movement and has much less correlation between days than price itself.

3.3.2 Digital News Data

To extract emotion and opinions of public towards a financial market, the corpus of textual data is built up with social media content. When choosing the source of social media, there are criteria that this work followed. Firstly, considering the importance of ensuring the selected source can serve as an accurate representation of public, formal national newspapers are preferred source in this study. Since the work is targeted at market, ideally the newspapers would be finance specific. Secondly, Chinese Natural Language processing techniques are less developed than it for English, the optimal data source should have English version published. Thirdly, the official data source should be open source and the historical data should be available to access.

Based on the criteria, the choices narrowed down (see Table 3.4). Among listed all sources, this work took Xinhua Finance News to be the source of textual data as it is the official channel to provide Chinas business news and financial information daily. Xinhua Finance News is targeted at financial news only and has focus on Chinese market. It also publishes in English officially and availability on LexisNexis, in a way it guarantees the content of the English text is correct and genuine.

Table 3.4: Sources of available newspapers in China collected from LexisNexis Business. Note that all publisher is based in China, however, they have different publishing frequency: mostly daily but ranging from 4-5 times a week to everyday.

Source	Coverage	Frequency	Publisher
China Daily	09/06/2010-*present*	Daily (Monday - Saturday)	China Daily Information
The China Post	11/21/2013-*present*	Daily (4-5 times a week)	China Post
Global Time (China)	08/09/2013-*present*	Daily (Monday-Friday)	Global Times
Xinhua Finance News	08/01/2015-*present*	Daily (Monday to Sunday)	Xinhua Finance Ahency (AP) Limited

News articles from Xinhua Finance News were collected and downloaded from LexisNexis Business, which is a news engine providing services, including query terms in publications and available to download. Access to LexisNexis Business engine services

were covered under an academic license obtained by Trinity College Dublin (TCD). There is an API provided by LexisNexis for retrieving data but only issued with a specific license which is not covered by TCD, so the collecting and downloading in this work was performed manually.

3.4 News Analysis

The purpose of sentiment analysis is to extract emotions and opinion from textual content, the phase of sentiment analysis of this work is to focus on identify and extract negativity shown in time series of the corpus and later on to be aggregated with financial data for statistical analysis. The approach for this phase is introduced here.

3.4.1 Text Analytic

This study took bag-of-words model implementation for text analysis. The bag-of-words model is one of the most popular representation methods for object categorization which is adopted in text analysis as well. Natural language words have been categorized to represent various emotions or sentiment according to psychology researchers and general purpose dictionary has been proposed and widely adopted by studies in sentiment analysis.

The list of target words is generalised into dictionary, the word is evaluated from different contextual linguistic aspects: strong to weak (based on strength), active to passive (based on action) and negative to positive (based on evaluation) (see Figure 3.2). The dictionary contains 6 affect categories as the column and terms (sentiment tokens) as rows (see table 3.5).

.png

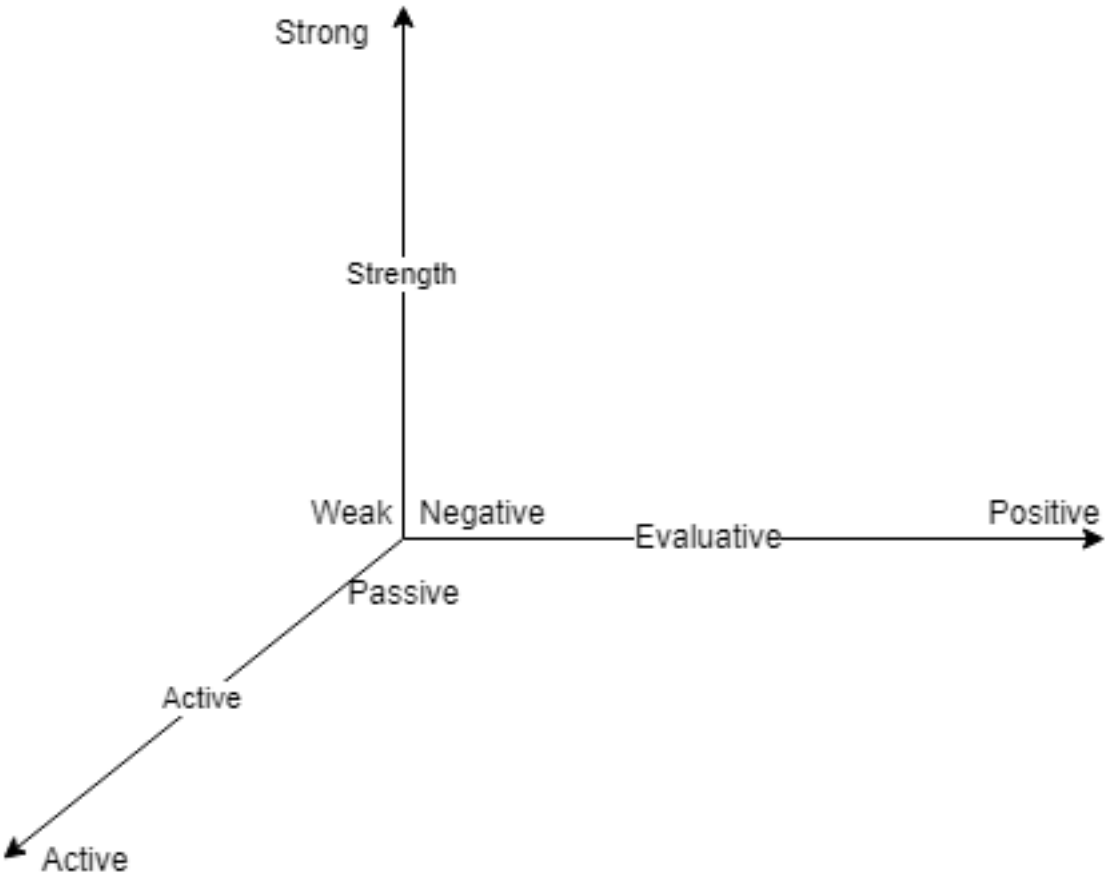

Figure 3.2: Word Affect

Table 3.5: Dictionary Example.

Term	Category							
	Strong Active	Negative	negative only	Active	...	Large_CN	Med_CN	Small_CN
cancel	-	Negative	negative only	Active				
prepare	Strong active	-	-	Active				
...								
360 Security Technology Inc.						Large_CN		
Anji Foodstuff co., ltd								Small_CN
...								

The dictionary contains k category including affect categories and three more custom categories for entity names, and k tokens: a token can belong to various categories and k should be much less than k (see Figure 3.2).

The key idea of the bag-of-words model is to take pieces of text: which represented as collection of words, and to tokenize them into collection of tokens. For each piece of texts, the distribution of each token in the text with be analysed and later the token will be matched against a list of target words: the list of target words can be affect words, for example, negative or positive (see Figure 3.5). With distribution of each token in the text and list of affect words, the affect score can be calculated for the text, and this study used list of negative words which gives out the negative sentiment score for textual data.

Assuming there are m days in total in the news time series and for each day, there are n pieces of text. For each text T, it parsed through a tokenisation process where text parsed with space and tokenised into unique words as token with frequency for the text. Next step is for each token, it matches through all the category in the dictionary, the category stands for the affect of the term. The algorithm takes a note of the frequency of each category in this text piece and it later be used in calculating the

score for the affect of the text. With the affect score of each text in the day, it sums up the affect score of the day, which builds up to the sentiment time series (see pseudo code below).

Algorithm 1 Calculate the affect score $CAT(K, D)$ for category k and day i

```
i ← 0
for i < m do
  j ← 0
  for j < n do
    Tokenize(T_j^(i) → {t_1, ..., t_l})
    k ← 0
    for k < category do
      k' ← 0
      for k' < length do
        k" ← 0
        for k'' < row do
          if t_k' = entry(k'') then
            CAT(k, i) ← CAT(k, i) + 1
          end if
          k'' ← k'' + 1
        end for
        write(k, CAT(k, i))
        k' ← k' + 1
      end for
      k ← k + 1
    end for
    j ← j + 1
  end for
  i ← i + 1
end for=0
```

3.4.2 Name Entity Dictionary

The basic dictionary employed by Rocksteady is sufficient for this study to use to analyse negativity from textual content. However, this study took a specific interest in how mentioning companies in the text could have an impact on price changes, which is to investigate whether oral sentiment has a correlation with price changes. For this

reason, a dictionary containing categories of companies in China was created in order to pick out certain terms and phrases as well during the sentiment analysis process.

There are over ten thousand companies exist in China, however, only less than two thousand of them are listed in Shanghai Stock Exchange. Since this work is focused on China market and taking SSE index as financial data, the dictionary only included the companies listed in SSE. Based on the total market capitalisation of each, companies were categorised into three: large, median and small. Also, for each company, the dictionary included its full name and short name. Adding the dictionary of companies name to text analytics system, it generates the time series of oral sentiment of company as well: the number of large, median or small company name mentioned in each text in the time series.

The companys name with be one of the token from text and when matching against the list of company, it takes a distribution in company density. The score of company density calculated by the model was taken into consideration in this study for analysing impact of companys density on price movement. With the added glossary, using bag-of-words model, the frequency of sentiment categories and frequency of entity on a daily basis will be detected and stored.

3.5 Statistical Modelling

3.5.1 Data Aggregation

The return series was generated by preprocessing methods, and a sentiment time series, including negative score and large/med/small company score, was extracted from Rocksteady after perform sentiment analysis; both of the time series were at daily-frequency. Two time series are combined based on date into one multivariate time series using python. A matching method was implemented to first look up the return by date which is in the sentiment series and added return into a new column in the sheet of sentiment series data. If there is no returns of the date then zero was input to prevent conflicts on missing data issue when later on modelling.

3.5.2 Vector Autoregression (VAR) Analysis

In pioneering and highly influential studies[18], Sims argues and presents the use of an econometric technique known as vector autoregression (VAR) for the analysis of multivariate time series which is widely adopted by studies ever since in the field of economics. VAR models consider multivariate time series as being composed of two types of variables: endogenous variables that are inherently internal in nature to the system, and exogenous variables that are viewed as external to the system.

This study is taking use of vector autoregressive techniques for analysing aggregated multivariate time series. As the aim of the study is to examine the impact of sentiment and company density on financial market, here financial return is considered as the endogenous variable which is on the lefthand side of the model and it also serves as proxy for financial market performance; negative sentiment and oral sentiments of company are considered as exogenous variables of the system as they are regarded as an external factor affecting return. In additional, extra variables for key events ($Crash1415, TrumpTalk, Congress$) are created to account for effect posed to return and included as exogenous variable of the system (see Chapter 4).

Evaluation of the model is done through examination of residual sum squared (RSS) and goodness of fit term (R^2) of each model. When comparing two models, the one with higher R^2 value and lower RSS value suggests that this model has a better fit for the data set where the accuracy is higher. Model estimation takes use of the terms and evaluate if the model is better by introducing now variables or not and then decide to take it further or not.

The GNU Regression, Econometrics and Time series Library (GRETL) was used to perform VAR analyses. The generalised VAR model for returns with lagged terms is:

$$r_t = const. + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \alpha_3 r_{t-3} + ... + \alpha_n r_{t-n} + \varepsilon_t \tag{3.3}$$

where r_t is return at time t, ε_t is residual term at time t and α_{t-i} is the weight of returns over the period of time from $(t-i)$ to t. As this study takes five lags of return as there are five trading days in a week, which is the first model present here:

$$r_t = const. + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \alpha_3 r_{t-3} + \alpha_4 r_{t-4} + \alpha_5 r_{t-5} + \varepsilon_t \tag{3.4}$$

The equation can be simplified by using an expression to replace all of the lag terms (Model 1):

$$r_t = const. + \alpha L_5 r_5 + \varepsilon_t \tag{3.5}$$

where

$$\alpha L_5 r_5 = \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \alpha_3 r_{t-3} + \alpha_4 r_{t-4} + \alpha_5 r_{t-5}$$

The first model is improved by adding variable to take off the possible impacts on return during the periods directly after events (national party congress (P), international events such as Trumps speech (T) and stock crash period from 2014 to 2015 (C)):

$$r_t = const. + \alpha L_5 r_5 + \beta C_t + \gamma P_t + \delta T_t + \varepsilon_t \tag{3.6}$$

where , , are the weight coefficients of proxy variables representing periods of market crash (C), national party congress (P) and Trumps public speech mentioning China (T) respectively.

The second VAR model introduced five lags of the negative sentiment proxy:

$$r_t = const. + \alpha L_5 r_5 + \beta C_t + \gamma P_t + \delta T_t + \theta L_5 S_5 + \varepsilon_t \tag{3.7}$$

where θ was the weight coefficient of proxy variable representing negative sentiment.

The third VAR model based on the second one, introducing five lags of company density proxy. The company are categorised into three: large, median and small. The new model was represented as shown below:

$$r_t = const. + \alpha L_5 r_5 + \beta C_t + \gamma P_t + \delta T_t + \theta L_t + \rho L_5 Lar_5 + \sigma L_5 Med_5 + \tau L_5 Sml_5 + \varepsilon_t \tag{3.8}$$

where ρ, σ, τ were the weight coefficient of variables representing the density of large Chinese company, median Chinese company and small Chinese company respectively.

For each model, this study examines residual sum squared (RSS) and goodness of fit term (R^2) and with further investigation into the coefficient as in if there is a significance shown towards the return. Eventually, this study would have evaluate the impact of each of these variables (negative sentiment, oral sentiments of company) as influencing factors on return.

3.6 Conclusion

Here, the methods this study took is introduced, including data collection and processing, text analysis and statistical modelling. How method has been implemented in the study will later be discussed in details with case study. Case study done with focus on Chinese market, taking Shanghai Stock Exchange as financial data source, with newspaper pieces from Xinhua Finance as sentiment source, and results from analysing aggregated multivariate modelling will be discussed thoroughly for three models introduced previously next.

Chapter 4

Case Study and Results

4.1 Introduction

4.1.1 Financial Market

Introduction

Recall that the index is an aggregate value of all listed stocks in the SSE including A shares and B shares: A shares are traded in Chinese Yuan (CNY) while B shares traded in US dollar; each listed stocks share price at time t referred to as $p(t)$, is multiplied by the total number of shares issued and sum that production of all stocks up to total market capitalisation. For calculating the SSE composite index, B shares stocks are denominated in US dollar.

The index values are calculated based on the market which represents the market only for that time; the economic and global environment is not still and the scale of measure may change every single second, for example, currency value changes due to inflation and other economic factors on almost a monthly basis we cannot compare the values directly at two different times as the difference may not be how it seems to be in reality. Furthermore, the index is heavily correlated one day lag correlation is 99.2% and so on (see Table 3.1 for details).

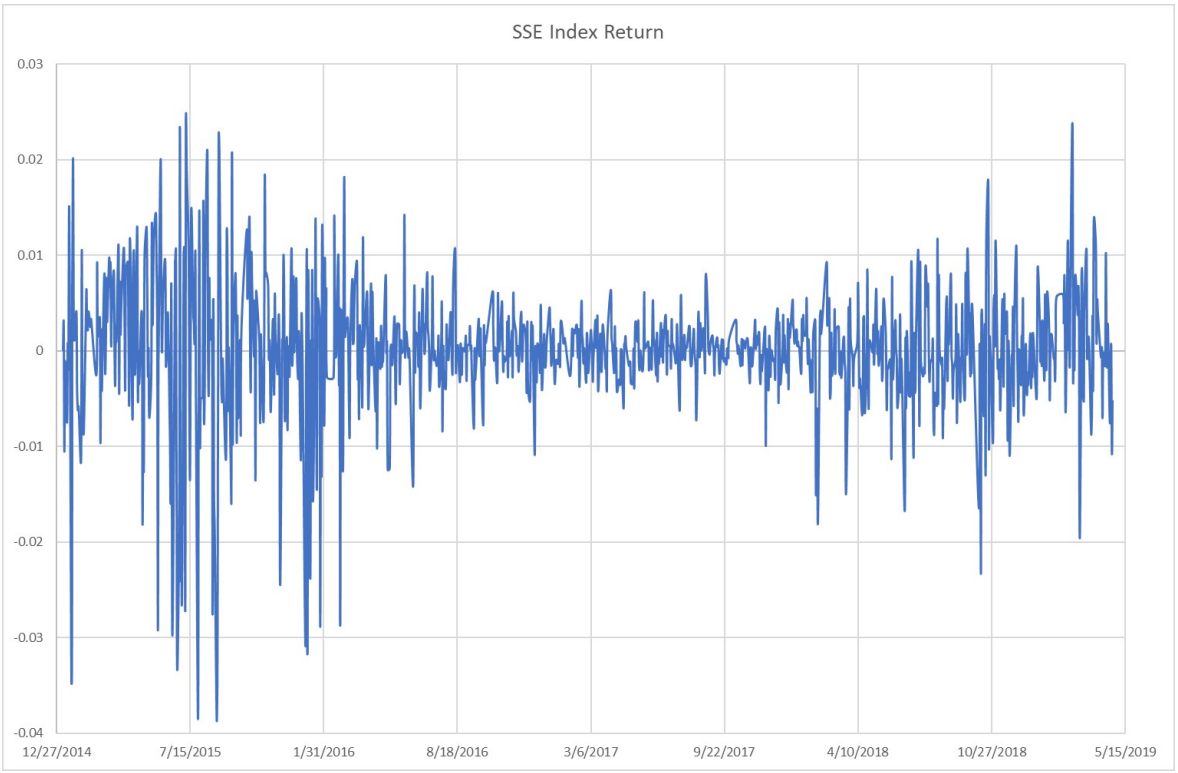

Figure 4.1: Four years of Shanghai Stock Exchange composite index daily returns. It shows different volatility in Shanghai Stock Exchange index returns: for example, around 15th July in 2015, the SEE index return shows a high volatility cluster, while low volatility only contains few, if any, large positive or negative return: for example, from June to September in 2015, the SEE index return shows a low volatility cluster.

The return values appear to have many more fluctuations than the price series and they are. The return values for the SSE index show a series which looks noise but fulfills an important condition for a perfect market: markets help in the discovery of prices. The fluctuations of the SSE return series are higher than for the price series (see Figure 1.1, Figure 4.1). The return series is far less correlated then index series one day lag correlation is only 5.25% and so on (see Table 3.3 for details).

An empirical phenomenon volatility clustering was discovered first by Benoit Mandelbrot that of price, large changes tend to be followed by large changes of either sign and small changes tend to be followed by small changes [19]. In the time series, markets show different volatility in different periods of time; high volatility cluster contains

several large positive return and large negative return (see Figure 4.1).

Stylised Facts

There are stylised facts for financial returns which are three general properties that can be seen in any set of returns [19].

Table 4.1: Descriptive Analysis on Return Series of Shanghai Stock Exchange Composite Index. Note that it is not normal distribution. Negative skewness (-1.19) indicates that the left tail is stronger and the mass of the distribution is concentrated on the right-hand side of the figure (see Table 4); the kurtosis of any normal distribution is 3 and it is above 6 in the case of SEE index return, which indicates that it has fatter tail than a normal distribution

Mean	0.00%
Standard Error	0.0002
Median	0.0003
Mode	N/A
Standard Deviation	0.007
Sample Variance	0.00005
Kurtosis	6.52
Skewness	-1.19
Range	0.06
Minimum	-0.04
Maximum	0.02
Sum	-0.04
Count	1051
Confidence Level (95.0%)	0.0004

First of all, distribution of return is not normal. It shows the same in SEE index data, the distribution of SSE index returns is approximately symmetric; it has fat tails, until nearly four standard deviation from mean, and a high peak lies between 0 and 0.5 standard deviation from mean (see Figure 4.2).

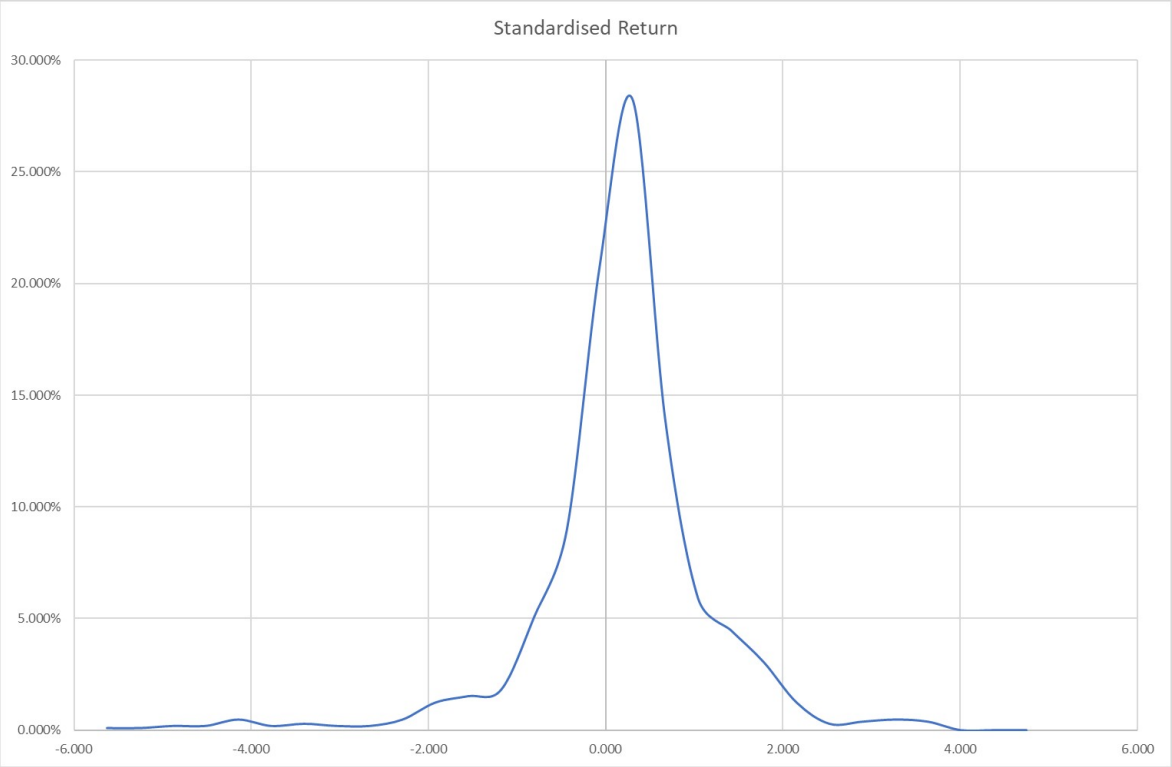

Figure 4.2: SSE composite returns distribution.

Skewness statistics are used to represent the symmetry of distributions and kurtosis statistics are used to interpret as a measure of similarity of the distribution to a normal distribution. The SSE index spot series has skewness equal to -1.187 for four years from 2015 to 2019, and its kurtosis is 6.516. Negative skewness indicates that the left tail is stronger and the mass of the distribution is concentrated on the right-hand side of the figure (see Table4.2); the kurtosis of any normal distribution is 3 and it is above 6 in the case of SEE index return, which indicates that it has fatter tail than a normal distribution (see table 4.3). SEE index return distribution is not normal and has stronger tail in the left indicates that there are more negative returns than positive returns, which means there appears more extreme price falling then rising in the series. Relative frequency of Shanghai Stock Exchange composite index returns during the period of time from 2015 to 2019.

Let r to be the return value, $\bar{r}$ to be the mean of daily return in the series, s to be the standard deviation of the return series. The frequency of observed daily returns in

Table 4.2: Relative frequency of Shanghai Stock Exchange composite index returns during the period of time from 2015 to 2019.

	Percentage of return within/beyond the number of standard deviations (s) from the mean ($\bar{r}$) beyond							
	within		beyond					
	0.25	0.5	1	1.5	2	3	4	5
Normal Distribution	19.74	38.29	31.73	13.36	4.55	0.27	0.01	0.00
SSE index	20.742	57.945	23.122	10.943	5.139	2.57	0.00	0.00

the range from $\bar{r}-0.5$ to $\bar{r}+0.5$ is higher than it expected to be in a normal distribution, meaning the high peak is lower than it of a normal distribution; the frequency in the range from $\bar{r}+0.25$ to $\bar{r}+0.5$ is higher than it in the rage from $\bar{r}$ to $\bar{r}-0.25$, meaning the high peak of return distribution is slightly away from the middle (see Table 4.3).

Table 4.3: Frequency for standardised daily returns of Shanghai Stock Exchange composite index. The frequency of observations lies below $\bar{r}-3r$ or above $\bar{r}+3s$ is approximate seven times as it of normal distribution, corresponding to the high value of kurtosis.

Range	Observed	Normal	Observed minus normal
0 to 0.25	20.742%	19.7%	1.042%
0.25 to 0.5	37.203%	18.6%	18.603%
0.5 to 1	18.935%	30.0%	-11.065%
1.0 to 1.5	12.179%	18.4%	-6.221%
1.5 to 2	5.804%	8.8%	-2.996%
2 to 3	2.569%	4.3%	-1.731%
3+	2.57%	0.3%	2.27%

Moving on to the dependence between the return for time period, the correlation between returns for time periods t and $t-\tau$, let τ be the interval of time period. The correlation between returns τ periods apart is then estimated from n observations in total by the *sample autocorrelation* at lag τ , with $\bar{r}$ the sample mean of all observed

return (see Equation 4.1).

$$\hat{\rho}_{r,\tau} = (\sum_{t=1}^{n-\tau}(r_t - \bar{r})(r_{t+\tau} - \bar{r}))/(\sum_{t=1}^{n}(r_t - \bar{r})^2), \tau > 0 \tag{4.1}$$

Second important stylised fact for daily return is that there is almost no correlation between returns for different days. So is the case with SEE index, the autocorrelations of SEE index returns are generally close to zero, and value varies from positive to negative with lag from 1 to 5 (see Table 3.3).

When focusing on one day lag autocorrelation for the entire return series: time period from t to $t+1$, there is no clear linear dependence shown between t and $t+1$, and it is hard to make out any nonlinear dependence from it as well:as more than 95% of data points lying in the centre $(0, 0)$ of the plot. There is no clear observation in whether $p(t+1)$ is positive or not depending on the positivity of $p(t)$:there are points that have negative $p(t)$ and positive $p(t+1)$ or negative $p(t+1)$ (see Figure 4.3 for details).

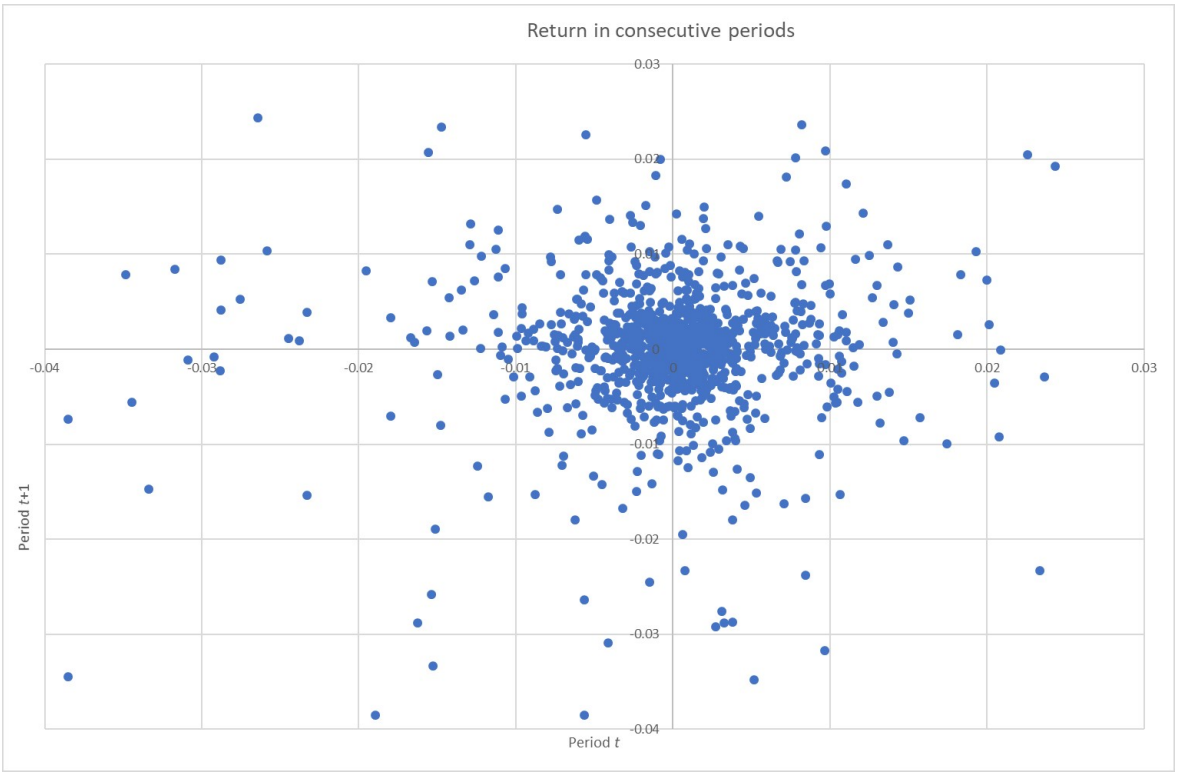

Figure 4.3: Shanghai Stock Exchange composite index returns in consecutive periods. X-axis to be the $p(t)$ while y-axis to be the $p(t+1)$, the day and the day after. Most of the points lie in the overlapping area of negative part of x-axis and y-axis. While points with same value of $p(t)$ have random different value of $p(t)$ from observation.

However, function of returns can have substantial autocorrelations even though returns do not. Third major stylised fact goes, there is positive dependence between absolute returns on nearby days, and likewise for squared returns.

The autocorrelation between $|r_t|$ and $|r_{t+1}|$ is 0.284 and there appears no clear observation of dependence between variables (see figure 4.4). However, as increases, there is more chance of a high value of $|r_{t+1}|$.

The autocorrelations of absolute values of returns are always positive at a lag of one day and the positive dependence continuing to be found for further lags. Power transformation of absolute returns: for example, squared returns , shows positive dependence but to a lesser degree comparing to absolute returns (see Table 4.4 and Table 4.5).

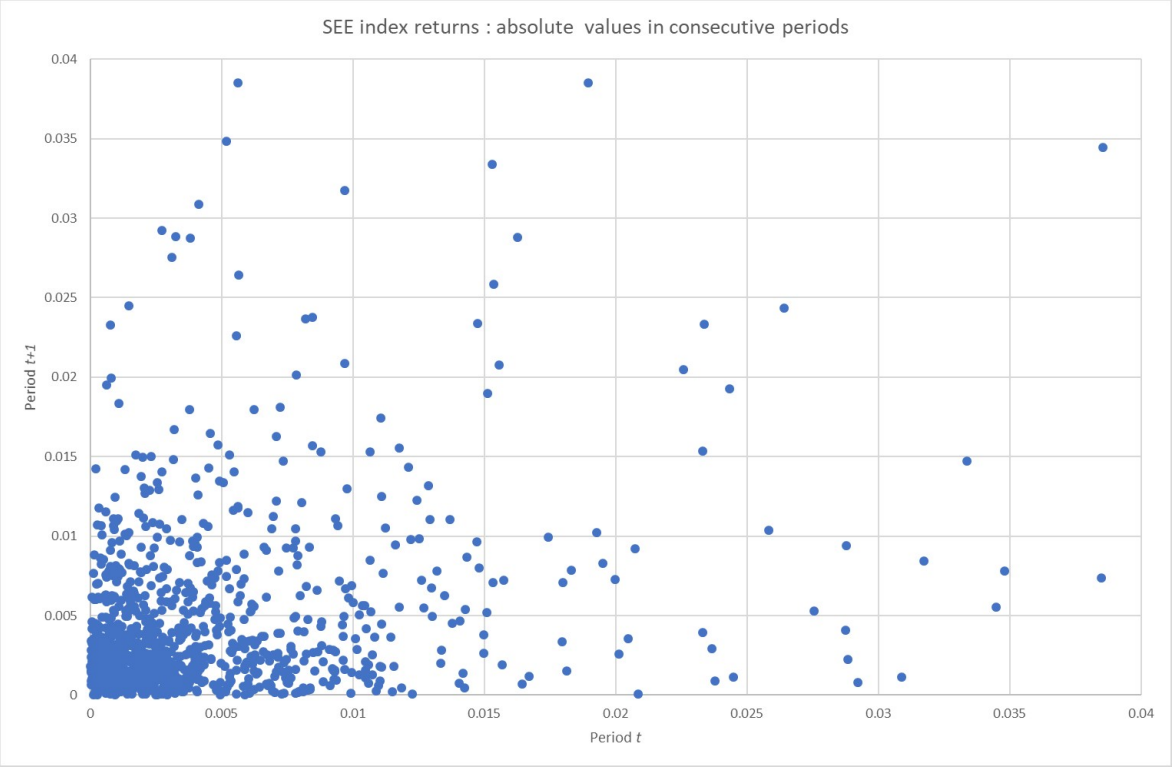

Figure 4.4: Shanghai Stock Exchange composite index returns: absolute values in consecutive periods. X-axis to be the p(t) while y-axis to be the p(t+1), the day and the day after.

Table 4.4: Autocorrelations for absolute value of Shanghai Stock Exchange composite index returns. Note that correlation of absolute return is higher than it of original return value. Recall the correlation of prices return with one day lag is 5.25% it is 1.07% with five days of lag, and the correlation of prices with one day lag is 99.2% and for five days of lag it is 95.75%, the correlation between absolute return and it of origin value of return are both lower than it of price, however, it of return itself is the lowest.

Lag(s)	**1**	**2**	**3**	**4**	**5**
Correlation	28.4445%	34.3761%	33.0128%	29.1966%	29.3344%

Table 4.5: Autocorrelations for squared absolute value of Shanghai Stock Exchange composite returns. Recall the correlation of prices return with one day lag is 5.25% it is 1.07% with five days of lag, and the correlation of prices with one day lag is 99.2% and for five days of lag it is 95.75%, the correlation of squared absolute return, absolute return and origin value of return are lower than it of price, however, it of return is the lowest.

Lag(s)	**1**	**2**	**3**	**4**	**5**
Correlation	22.86%	27.06%	26.50%	23.36%	20.47%

The fact that high dependence in series of absolute return proves the point that returns are not made up of independent and identically distributed (i.d.d.) random variables. Based on the volatility clustering effect found in return series, the observed dependence can be explained simply as a reflection on changes in the flow of relevant information to the market which is worth further investigation.

4.1.2 News about the Market

News Source

Xinhua Finance News is chosen to be the source of textual data in this study as it is the official channel to provide Chinas business news and financial information daily. Xinhua Finance News, provided by Xinhua Finance Agency, is targeted at financial news only and has focus on Chinese market.

As this study is interested in Chinese market, in order to narrow down the text retrieved and for the articles to be more financial regarded, terms *economy* and *economics*

are used in query (see Table 4.6).

Table 4.6: Search Query Strategy

Market	Query Term(s)	Criteria
Chinese Market	"economics", "economy", "economy*"	Group high similarity, major mentioned

Stylised Facts

Xinhua Finance News pieces are collected for the period of time from January 2015 to May 2019 with a total number of 18181 articles, 16823 valid articles excluding articles less than 140 words. The corpus covers 1151 days with an average of 6 articles per day. Text Analysis are on categories specifically, this study takes negative sentiment, large company density, median company density and small company density affect.

Rocksteady was used to perform text analysis on the corpus created. As this study was to perform time series analysis, news pieces in the corpus were grouped by day in order to generate the corresponding sentiment time series of negative sentiment and oral sentiments of small company, median company and large company. (see Figure 4.5,4.6,4.7,4.8 for details).

Table 4.7: Descriptive Analysis for company name entity and sentiment proxy extracted from text.

	Large_CN	Med_CN	Small_CN	Sentiment
Mean	0.23	0.002	4E-05	1.4
Standard Error	0.02	0.001	5E-05	0.05
Median	0.04	0	0	1.3
Standard Deviation	0.4	0.02	0.001	1.1
Sample Variance	0.2	0.0004	1E-06	1.1
Kurtosis	15.5	301.7	493	1.6
Skewness	3.4	16.5	22	1.1
Range	3.4	0.4	0.02	6
Minimum	0	0	0	0
Maximum	3.4	0.4	0.02	6
Sum	114.5	0.8	0.02	704
Count	493	493	493	493

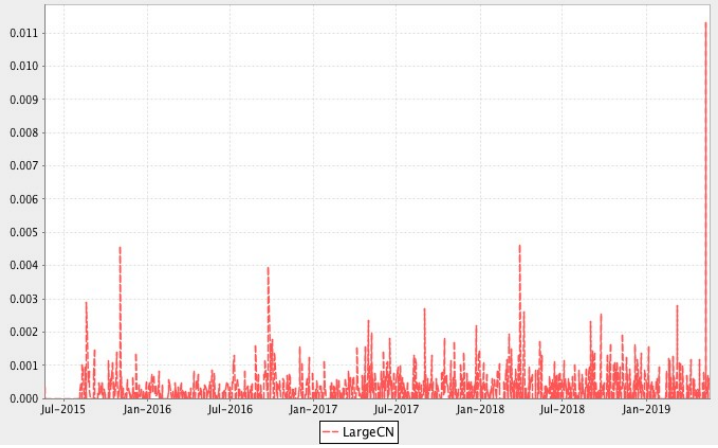

Figure 4.5: Large Chinese Company Density affect in News Corpus in percentage. $Large\_CN$ series generated with percentage of the affect calculated, shown at a daily basis. Note that the scale is from 0 to 1.1%.

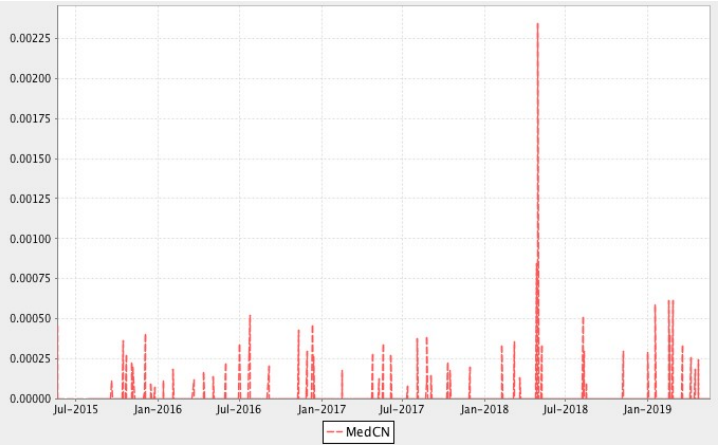

Figure 4.6: Median Chinese Company Density affect in News Corpus in percentage. Med_CN series generated with percentage of the affect calculated, shown at a daily basis. Note that the scale is from 0 to 0.225%.

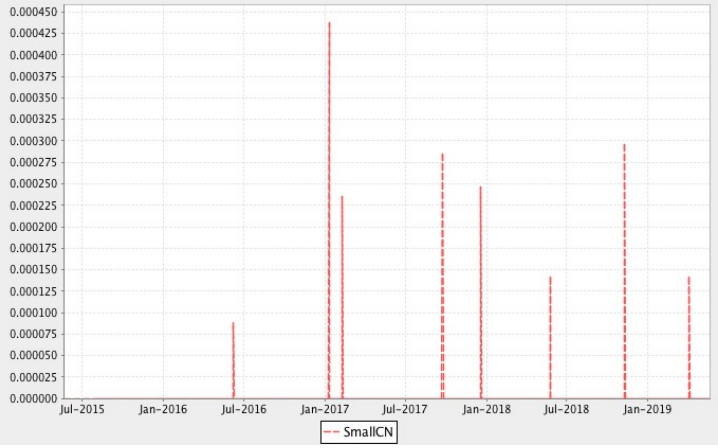

Figure 4.7: Small Chinese Company Density Affect in News Corpus in percentage. *Small_CN* series generated with percentage of the affect calculated, shown at a daily basis. Note that the scale is from 0 to 0.045%.

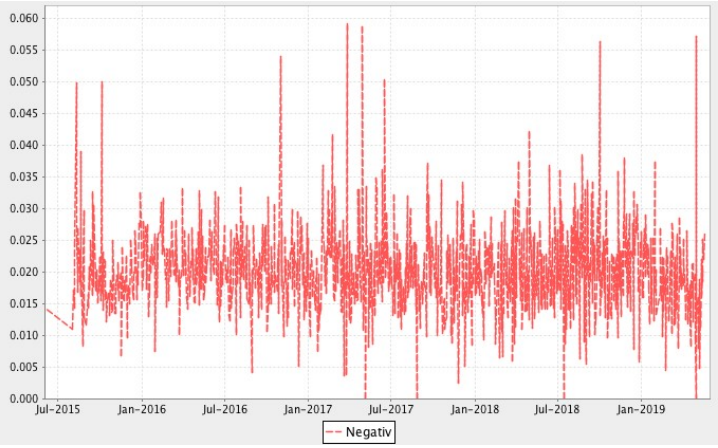

Figure 4.8: Sentiment Affect in News Corpus in Percentage. Negative sentiment series generated with percentage of the affect calculated, shown at a daily basis. Note that the scale is from 0 to 6.0%.

4.2 Financial Time Series Analysis

This study in order to find a better fit statistical model for the market in general, took key events for abnormal behaviours in the market and introduce variables to

balance the uncertainty they introduced to the market while modelling. The variable introduced including crash period, trump speech and national party congress period.

As discussed previously, there was a boom in the market where the price kept rising from average price to almost double of it; and right after it reached its peak it fell continuously back to average. The market was not behaving as price should be during the time which is also known as crash period. For modelling purpose, variable regarding the misbehaviour is introduced as variable Crash1415 and dates within the period of time from January 8th to August 24th in 2015 are marked.

During Trumps first election campaign, he publically mentioned in his speech on China specifically a few times, and from the price series of SSE, downticks are observed after those dates. This study includes variable named as Trumptalk and it marks the dates where Trump gave speech on China specifically.

National congress of the communist party of china is a party congress that is held every five years. The 19th national congress was held between 18 and 24 October 2017 which is the latest one. The national congress is where the committee come together, talk and make announcement regarding entire nation; and possibly new ideology would be written into the partys constitution. A national event as such is the focus of the nation by the time and such dates have effects on price the similar mechanism as calendar affect. Variable named NationalCongress is introduced in the modelling which marks the date that 19th national congress of communist party took place.

Table 4.8: Panel regression model on return with lags of five-days: comparison shown with variables introduced and without. Model 1a (1b, 1c, 1d, 1e) is return with one (to five) day(s) lag while model 2a (2b, 2c, 2d, 2e) include variables: crash period, Trump speech dates and national party meeting dates. Note that upon introductions of crash variable, trump speech and national party meeting, the residual sum squared (RSS) for model with lag of day(s) from 1 to 5 shows slight decrease and shows increase in each model: for example, RSS and are 0.0255 and 0.0186 for model 1e while that are 0.0253 and 0.0226 for model 2e. Also, note that the return with five days lags carries a significance.

	Model 1a	Model 2a	Model 1b	Model 2b	Model 1c	Model 2c	Model 1d	Model 2d	Model 1e	Model 2e
Const.	-1	-1	-1	-1	-1	0.9	-1	-1	-1	-1
r_{t-1}	-118	-116	-124	-123	-143	-150	-67	-76	-83	-86
r_{t-2}	-	-	254	267	257	271	299	303	332	337
r_{t-3}	-	-	-	-	746	789	746	776	708	751
r_{t-4}	-	-	-	-	-	-	*-1022*	*-994*	*-1041*	*-992*
r_{t-5}	-	-	-	-	-	-	-	-	-457	-413
Crash1415	-	-11	-	-18	-	-23	-	-16	-	-31
TrumpTalk	-	-1		-0.5	-	-1	-	-6	-	-4
Party Congress	-	2	-	2	-	3	-	2	-	2
RSS	0.026	0.026	0.026	0.026	0.026	0.025	0.025	0.025	0.025	0.025
R^2	0.0001	0.0009	0.0008	0.0026	0.0063	0.0089	0.0167	0.0180	0.0186	0.0225

As for china market, it shows statistical significance of the return variable on the fourth lag term, also it should be noted that it contributes consistently to return relatively small, around 990 basis points.

Table 4.9: Panel regressions model on sentiment, small company density, median company density and large company density with no day lag.

	Model 3a	**Model 3b**	**Model 3c**	**Model 3d**
Const.	**−1**	−1	−1	−1
r_{t-1}	**−99**	−89	−87	−80
r_{t-2}	**334**	336	337	316
r_{t-3}	**745**	751	752	750
r_{t-4}	***−985***	*−991*	*−991*	*−988*
r_{t-5}	**−404**	−415	−414	−407
Sentiment	-	−0.1	-	-
Small_CN	-	-	0.1	-
Med_CN	-	-	-	1.
Large_CN	−1	-	-	-
Crash1415	−30	−30	−31	−30
TrumpTalk	−4	−5	−4	−4
National Congress	3	2	2	3
RSS	0.025	0.025	0.025	0.025
R^2	0.0236	0.0225	0.0226	0.0232

4.2.1 Model Estimation

Please note that in all model estimation, coefficient terms are shown in basis point where one basis point is equal to 0.01%. Also for better view ability coefficient values are written in different font to show level of significance, ***Bold Italic*** for significance at more than 99%, **Bold** stands for significance at level of 95% and *Italic* is for 90% level.

4.2.2 Model Estimation with Variant

The relationships between the negative sentiments, entity variable generated from corpus of newspapers and the returns of SSE index are examined using a vector autoregression model (see chapter 3). The results for model estimation with each independent

variables with no day lagged show that 3-day lag of returns are statistically significant at 10% level with an 990 basis point impact on returns. But when examining the RSS and R^2, model 3b shows the lowest R^2 which indicates less variance in returns (see Table 2.).

From the model estimation, none of the variants holds statistical significance for coefficient to return over 10%, however, examinations on each variant with day lags could show significance as return with 4-day lags shows it at level more than 10% on return.

Model Estimation with Sentiment

Further examine on sentiment variable with 5-day lags shows that sentiment with 2-day lags is statistically significant to return at level more than 5% with a 5 basis point impact on return, which is comparably small. And 3-day lag sentiment variable shows statistical significance to return at level of 5 % with a 3 basis point. In addition, with more lagged sentiment introduced, R^2 increases which indicate a worse fit for all return data. Also, with introduction of lagged sentiment in the model, return with 4-day lag shows statistically significance to returns at more than 10% level with a 1000 basis point impact on returns.

Model Estimation with Company Density

As the company entity is categorised into three: small, median and large, models are built using each as independent variable.

The examination on small company density $Small\_CN$ shows that small company density with 1-day lags is statistically significant to return at level more than 1% with a 1 basis point impact on return, which is comparably small. Also, with introduction of lagged small company density in the model, return with 4-day lag shows statistically significance to returns at more than 10% level with a 992 basis point impact on returns.

Table 4.10: Panel regression model with small Chinese company frequency $Small\_CN$ variable. Note that the four day lagged return carries significance to the returns and it is consistently approximate 992 basis point (in consistency with model 2). The one-day lag of $Small\_CN$ carries significance in return consistently but relatively small, approximate 1.22 basis points. Residual sum squared shows a slight decrease with more day lags introduced to the model while shows relatively considerable increase.

	Model 4a	**Model 4b**	**Model 4c**	**Model 4d**	**Model 4e**
Const.	−1	−1	−1	−1	−1
r_{t-1}	−86	−88	−89	−89	−89
r_{t-2}	340	341	340	339	340
r_{t-3}	752	754	754	753	753
r_{t-4}	*−993*	*−993*	*−992*	*−992*	*−992*
r_{t-5}	−414	−415	−415	−414	−414
$Small\_CN_{t-1}$	***−1***	***−1***	***−1***	***−1***	***−1***
$Small\_CN_{t-2}$	-	−0.5	−0.5	−0.5	−0.5
$Small\_CN_{t-3}$	-	-	−0.1	−0.1	−0.1
$Small\_CN_{t-4}$	-	-	-	−0.2	−0.2
$Small\_CN_{t-5}$	-	-	-	-	0.2
Crash1415	−31	−31	−31	−31	−31
Trumptalk	−4	−4	−4	−4	−4
National Congress	2	2	2	2	2
RSS	0.025	0.025	0.025	0.025	0.025
R^2	0.0231	0.0232	0.0232	0.0232	0.0232

Examination on median company density Med_CN variable shows that median company density with 1-day lags is statistically significant to return at level more than 5% with a 2 basis point impact on return. And 2-day lag median company density variable shows statistical significance to return at level of 10 % with a 2 basis point. In addition, with more lagged median company density introduced, R^2 increases which indicate a worse fit for all return data. Also, with introduction of lagged sentiment in the model, return with 4-day lag shows statistically significance to returns at more than 10% level with a 1020 basis point impact on returns.

Table 4.11: Panel regression model with Med_CN variable introduced with lag of five-days. Note that with Med_CN variable introduced, fourth day lag return shows significance and the values are consistently around 1020 basis points. Also, significance found in one and two days lagged Med_CN values, both relatively small and consistently: around 2.6 basis points for one day lagged value and 2.0 basis points for two days lagged value.

	Model 5a	**Model 5b**	**Model 5c**	**Model 5d**	**Model 5e**
Const.	−1	−1	−1	−1	−1
r_{t-1}	−101	−120	−112	−111	−103
r_{t-2}	349	337	348	346	336
r_{t-3}	710	720	727	723	735
r_{t-4}	*−993*	**−1022**	**−1027**	**−1030**	*−1010*
r_{t-5}	−409	−412	−397	−395	−381
Med_CN_{t-1}	**2**	**2**	**2**	**2**	**2**
Med_CN_{t-2}	-	*2*	*2*	*2*	*2*
Med_CN_{t-3}	-	-	−1	−1	−1
Med_CN_{t-4}	-	-	-	0.3	0.3
Med_CN_{t-5}	-	-	-	-	−1
Crash1415	−30	−29	−30	−30	−29
Trumptalk	−4	−4	−4	−4	−4
National Congress	3	2	2	2	3
RSS	0.025	0.025	0.025	0.025	0.025
R^2	0.0252	0.0266	0.0270	0.0270	0.0283

Examine on large company density $Large\_CN$ variable shows no statistically significant to return. However, with more lagged sentiment introduced, R^2 increases which indicate a worse fit for all return data. Also, with introduction of lagged sentiment in the model, return with 4-day lag shows statistically significance to returns at more than 10% level with a 1000 basis point impact on returns.

Table 4.12: Panel regression model with $Large\_CN$ variable introduced with lag of five-days. Note that four days lagged return value carries significance and it is consistent above 900 basis points, decreases slightly with more lagged $Large\_CN$ introducing to the model. However, no significance observed in $Large\_CN$.

	Model 6a	**Model 6b**	**Model 6c**	**Model 6d**	**Model 6e**
Const.	−1	−1	−1	−1	−1
r_{t-1}	−85	−85	−85	−89	−84
r_{t-2}	338	340	339	336	339
r_{t-3}	752	753	759	760	762
r_{t-4}	*−991*	*−991*	*−985*	*−974*	*−975*
r_{t-5}	−414	−414	−412	−403	−408
$Large\_CN_{t-1}$	0.2	0.1	0.2	0.1	0.2
$Large\_CN_{t-2}$	-	0.1	0.09	0.1	0.1
$Large\_CN_{t-3}$	-	-	0.8	0.7	0.6
$Large\_CN_{t-4}$	-	-	-	1	1
$Large\_CN_{t-5}$	-	-	-	-	−0.8
Crash1415	−31	−31	−31	−31	−31
Trumptalk	−4	−4	−4	−4	−4
National Congress	2	2	1	0.3	0.9
RSS	0.025	0.025	0.025	0.025	0.025
R^2	0.0226	0.0226	0.0228	0.0235	0.0237

4.3 Panel Estimation

With previous estimation on model with single variable and return, further examinations were done introducing more than one variable to the model. As also seen previously, density of large company does not have a statistical significance on return. With the introduction of density of large company variable, median company variable shows that median company density with 1-day lags is statistically significant to return at level more than 10% with a 2-basis point impact on return while it is more than 5% without large company variable. And 2-day lag median company density variable

shows statistical significance to return at level of 10 % with a 2-basis point. However, the introduction of large company variable has no impact on the significance from sentiment and small company variable on returns: the examination on small company density is consistent with $Small\_CN$ alone which shows that small company density with 1-day lags is statistically significant to return at level more than 1% with a 1 basis point impact on return, which is comparably small; the examination on sentiment shows that sentiment with 2-day lags is statistically significant to return at level more than 5% with a 5 basis point impact on return, which is comparably small. And 3-day lag sentiment variable shows statistical significance to return at level of 5 % with a 3 basis point. In addition, with more lagged sentiment introduced, R^2 increases which indicate a worse fit for all return data.

Also, with introduction of large company density and one other variable in the model, return with 4-day lag shows statistically significance to returns at more than 10% level with a 970-basis point impact on returns.

From statistical model with median company density alone, the significance is the highest among all models: significance found in one and two days lagged Med_CN values, both relatively small but consistent: around 2.6 basic points for one day lagged value and 2.0 basic points for two days lagged value. Among models that include lagged return value, median company density and one other variant: small company density, large company density or sentiment (see Model 7e, 8e and 7f respectively), evaluated by RSS and R^2 metrics of each, model with median company density and small company density shows the best fit for return values. Additionally, median company density shows consistent but small statistical significance at level above 5% with a positive 2-basis point impact on return in the models mentioned above.

As for small company density, it shows that small company density with 1-day lags is statistically significant to return at level more than 1% with a negative 1-basis point impact on return, which is comparably small. However, when adding lagged sentiment as another variant into the model, fourth lagged term of small company density shows statistical significance to return at level more than 5% with a negative 0.4-basis point impact on return; and second lagged term shows significance at level above 10% with a negative 0.6-basis point impact on return as well, which were not found in model with only lagged return and small company density.

For negative sentiment proxy, the second lagged term shows a consistent but small

significance at level more than 5% with a negative 5-basis point impact on return, which is comparably small. And 3-day lag sentiment variable shows statistical significance to return at level of 5 % with a positive 3-basis point. Additionally, models with sentiment and small company density as variables showed that fourth lagged term of small company density having a statistical significance to return at level more than 5% with a negative 0.4-basis point impact on return while it doesnt show without sentiment as variant.

The significances shown from lagged variables only suggest that they have impact on returns to various extends and when it comes to modelling, the metrics of RSS and R^2 is of most concerned as it indicated the fitting of model to the dataset.

Although the model does not necessarily show an increase in better fitting with more variants introduced into it. By examining the RSS and R^2 metrics, model with langged return, small company density, median company density, large company density and sentiment was the best fit model among all models built in this study, with a 0.0435 as R^2.

Table 4.13: Panel regression model on return with sentiment variable introduced with lag of five-days. Note that four days lagged return value carries significance and it is consistent above 950 basic points. Also, sentiment variable with two and three days lags carry significance; the values are consistent and relative small: one day lagged values are approximate 5 basic point and two days lagged values are around 3 basic points.

	M 8e	**M 8f**	**M 8g**	**M 7e**	**M 7f**	**M 8**	**M 9**	**M 10**	**M 11**	**M 12**	**M 13**
Const.	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
r_{t-1}	−97	−87	−9	−105	−30	−13	−99	−24	−11	−33	−27
r_{t-2}	340	342	362	339	360	362	342	362	363	362	364
r_{t-3}	747	764	668	737	646	657	748	657	667	646	657
r_{t-4}	*−997*	*−976*	*−942*	*−1009*	*−979*	−959	*−997*	*−967*	*−943*	−980	*−968*
r_{t-5}	−379	−410	−412	−381	−385	−418	−380	−383	−413	−385	−384
Lrg_CN_{t-1}	0.4	0.2	0.3	-	-	-	0.4	0.6	0.3	-	0.
Lrg_CN_{t-2}	0.1	0.1	0.07	-	-	-	0.1	0.09	0.06	-	0.07
Lrg_CN_{t-3}	0.5	0.6	0.6	-	-	-	0.5	0.5	0.6	-	0.5
Lrg_CN_{t-4}	1	1	1	-	-	-	1	1	1	-	1
Lrg_CN_{t-5}	−1	−0.8	−0.5	-	-	-	−1	−0.6	−0.5	-	−0.6
Med_CN_{t-1}	*2*	-	-	**2**	*2*	-	*2*	*2*	-	*2*	*2*
Med_CN_{t-2}	*2*	-	-	*2*	**2**	-	*2*	**2**	-	**2**	**2**
Med_CN_{t-3}	−1	-	-	−1	−1	-	−1	−1	-	−1	−1
Med_CN_{t-4}	0.4	-	-	0.3	0.1	-	0.4	0.2	-	0.1	0.2
Med_CN_{t-5}	−1	-	-	−1	−1	-	−1	−1	-	−1	−1
Sml_CN_{t-1}	-	***−1***	-	***−1***	-	***−1***	***−1***	-	***−1***	***−1***	***−1***
Sml_CN_{t-2}	-	−0.5	-	−0.5	-	*−0.6*	−0.5	-	*−0.6*	*−0.6*	−0.5
Sml_CN_{t-3}	-	−0.1	-	−0.1	-	−0.09	−0.1	-	−0.09	−0.07	−0.07
Sml_CN_{t-4}	-	−0.1	-	−0.2	-	**−0.4**	−0.1	-	**−0.4**	**−0.4**	**−0.4**
Sml_CN_{t-5}	-	0.1	-	*0.2*	-	0.2	0.2	-	0.2	*0.3*	0.3
S_{t-1}	-	-	1	-	1	1	-	1	1	1	1
S_{t-2}	-	-	***−5***	-	***−5***	***−5***	-	***−5***	***−5***	***−5***	***−5***
S_{t-3}	-	-	**3**	-	**3**	**3**	-	**3**	**3**	**3**	**3**
S_{t-4}	-	-	0.3	-	0. 4	0.3	-	0.4	0.3	0.4	0.4
S_{t-5}	-	-	−1	-	−1	−1	-	−1	−1	−1	−1
Crash1415	−30	−31	−30	−30	−29	−30	−30	−29	−30	−29	−29
TrumpTalk	−4	−4	−8	−4	−8	−8	−4	−8	−8	−9	−8
National Congress	1	0.9	1	3	3	3	1	1	1	3	1
RSS	0.025	0.025	0.024	0.025	0.024	0.024	0.025	0.024	0.024	0.024	0.024
R^2	0.029	0.024	0.037	0.029	0.042	0.037	0.030	0.043	0.038	0.0423	0.044

With further examination on RSS and R^2 performance evaluation metrics for each model, it is indicated that the best estimate model on return is model 13, which considered lagged return value, small company density, median company density, large company density and sentiment as factors influencing the market.

4.4 Conslusions

This study examined the impact of several variables including negative sentiment and company density. It shows consistency with prior study in the delayed effect on price movement from sentiment and interesting finding of oral sentiment has been observed at Chinese market.

The Chinese market exhibits interesting behaviour, with oral sentiments of small company and median company carrying statistical significance. It is seen that large company density has no statistical significance in the Chinese market. The oral sentiment of small Chinese company shows statistical significance on price as well, comparably small (between one and two basis points) but at a high level of over 95%.

A consistent statistical significance is carried by negative sentiment either as single variant or with other variants, and it carries a consistent small contribution in the market: the second lagged term negatively contributes 5 basis point while the third lagged term positively carries 3 basis point to the return.

Chapter 5

Conclusion and Future Work

The remainder of this chapter will discuss the conclusion and contributions of this study (Chapter 5.1). A discussion of the limitations of the work is the presented along with the potential future works (Section 5.2).

5.1 Conslusions

The main objective of this dissertation was to combine methods from the area of textual analysis and time series modelling to process and analyse the sentiment from news source text to determine a relationship with movements in assets traded on financial markets.

A method has been developed for analysing the impact of news on the price of a stock or the value of an index of stock prices. The term news here refers to the frequency of mentioning name entities in a market, and potential sentiment expressed about the entity. The method contains crashing a price series of price and the fusion with an equivalent time series of sentiment in the news and/or time series comprising the frequency by which name entity is cited in the news. For fusion I have used vector autoregression (VAR) which helps to examine the impact of news on the price movement, e.g., a share or index.

This dissertation took Chinese market as focus to investigate the impact of news on price movement which fill the gap in the literature of finance and computing. Apart from negative sentiment which has been proven to be related to price movement and

of interest for recent years, this dissertation explored with oral sentiments which is difficult to quantify in any sense. Proxies for oral sentiments were calculated based on the frequency of name entity mentioned in the news which shown interesting relation with price movements.

This dissertation started with introduction to the topic and background of the study. With the observation and explanation of the abnormal behaviour in the financial market, it leaded the study to reveal the main interest and focus which is the sentiments. With the focus on price movement in market and public sentiments, this study conducted a thorough literature review on existing works in the field of finance and computing. Including sentiment analysis techniques and behavioural finance. How sentiment being introduced into the field of finance was also covered.

After exploring and evaluating the existing methods of sentiment analysis, the methods employed by this study is introduced. How data source be picked out and collected were introduced and this study chose bag-of-words model for text analytics to extract valuable sentiment proxies, and vector auto regression model for statistical analysis: including four models which independently assess the impact of exogenous variables such as negative sentiment and density of company on financial returns.

Case study presented with study of Chinese stock market, a merging market with information hidden in price movement during the period of time from 2015 to present, the study took the return series of Shanghai Stock Exchange Composite Index as the financial time series to better examine its movement. Negative sentiment was extracted from approximately xxx pieces of newspapers articles about economics published between May 2015 and April 2019 by Xinhua Finance. Special dictionary build for Chinese company in order to produce proxies for company density, which was also included as a variant in the VAR model. The results of the analysis are included in the same chapter where they are discussed individually. Finally, this is followed by a more general discussion of the results.

The analysis found that sentiment within formal media, specifically newspaper, plays a small role in explaining returns of an asset. Additionally, it showed that the oral sentiments that were using frequency of name entity as sentiment proxies, for example, company, also has an explanatory power to returns. The study categorised the entity into three: large company, median company and small company, the interesting founding suggested that frequency of large company mentioned in formal newspaper

barely shows a significant impact on return while median and small company density showed a small but consistent statistical significance on returns.

5.2 Future Works

As how this study being conducted, there are few limitations which can be fulfilled in future works, instead of manual effort, it was necessary to have functions that could harvest and scrape data from online sources, perform data processing and aggregation, content analysis, and time series modelling.

As previously mentioned, high-frequency negative sentiment extracted from tweets was significant across many weeks at different levels, but the conditions in which it is significant are less than clear, besides the initial observations that it has the most significant impact during periods of consistent volatility. Future research will be carried out into the investigation of when and why high-frequency sentiment is significant, by further analysing its effect on shorter time periods and by varying the conditions. If sentiment is to be used as part of a high-frequency trading strategy it would need to be possible to determine when to take sentiment into account and how much influence it should have. Furthermore, this work focused mostly on a firm-specific analysis in the commercial airline industry. It would be interesting to determine whether the results are consistent for other firms in the commercial airline industry, and how much they might differ in another industry altogether.

In conclusion, the negative sentiment found in tweets is very appropriate for use in high-frequency trading strategies, as there is a very significant relationship between sentiment and returns of a financial asset but the circumstances under which it is significant need to be thoroughly examined before it can be fully exploited.

Bibliography

[1] E. F. Fama, "Two pillars of asset pricing," *American Economic Review*, vol. 104, no. 6, pp. 1467–85, 2014.

[2] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[3] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 14–23, 2014.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[5] J. R. Nofsinger, "Social mood and financial economics," *The Journal of Behavioral Finance*, vol. 6, no. 3, pp. 144–160, 2005.

[6] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise trader risk in financial markets," *Journal of political Economy*, vol. 98, no. 4, pp. 703–738, 1990.

[7] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter i hope it is not as bad as i fear," *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[8] G. Kling and L. Gao, "Chinese institutional investors sentiment," *Journal of International Financial Markets, Institutions and Money*, vol. 18, no. 4, pp. 374–387, 2008.

[9] B. M. Barber and T. Odean, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *The review of financial studies*, vol. 21, no. 2, pp. 785–818, 2007.

[10] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *The journal of Finance*, vol. 61, no. 4, pp. 1645–1680, 2006.

[11] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[12] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," *The Journal of Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.

[13] K. Ahmad, J. Han, E. Hutson, C. Kearney, and S. Liu, "Media-expressed negative tone and firm-level stock returns," *Journal of Corporate Finance*, vol. 37, pp. 152–172, 2016.

[14] A. Bhardwaj, Y. Narayan, M. Dutta, *et al.*, "Sentiment analysis for indian stock market prediction using sensex and nifty," *Procedia Computer Science*, vol. 70, pp. 85–91, 2015.

[15] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts.," in *DART@ AI* IA*, pp. 59–68, 2014.

[16] A. Moreno-Ortiz and J. Fernández-Cruz, "Identifying polarity in financial texts for sentiment analysis: a corpus-based approach," *Procedia-Social and Behavioral Sciences*, vol. 198, pp. 330–338, 2015.

[17] S. Kelly, *News, Sentiment and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets*. PhD thesis, Trinity College Dublin, 2016.

[18] C. A. Sims, "Macroeconomics and reality," *Econometrica: journal of the Econometric Society*, pp. 1–48, 1980.

[19] S. J. Taylor, *Asset price dynamics, volatility, and prediction.* Princeton university press, 2011.