Market Disequilibrium and the Impact of News Sentiment

A Computational Approach and its Application to Aggregate Market and Individual Firms

Zeyan Zhao

A dissertation submitted to the University of Dublin

for the Degree of Doctor of Philosophy

School of Computer Science and Statistics

Trinity College, the University of Dublin

December 2018

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed

Zeyan Zhao

Abstract

The market price of any stock or financial instrument listed on an exchange would be assumed to be observed price, though we believe there is always a true price (or fair market value) of any. The market value and market price are identical only under conditions of market efficiency, equilibrium, and rational expectations. According to the theory of efficient market, if the observed price of a stock is beyond the market value, the price is expected to drop and vice versa. This is consistent with one of the properties of stock returns - mean reversion. However, the market doesn't always reverse. Although the long-term mean of returns tends towards zero, if we look at the small proportion of the entire market movements, some show extreme up/downwards movements. We believe sentiment takes place if the market is not reversing. In this thesis, I have used a systemic approach with two parts of the basic regression models to explore market behaviour: first, I have applied parametric models to check whether investor sentiment exists at both market- and firm-level. Second, we focused on whether we can find a non-parametric method to visualise the relationship between return/residual and sentiment. There are mainly five case studies in this thesis. The first case study confirmed the relationship between the proxy of sentiment (extracted from the textual corpus) and DJIA market returns using linear regression models. The second and the third case studies used the same method to observe the relationship between sentiment and index returns in Danish and Chinese markets. The fourth study visualised the relationship between the proxy of sentiment and the return residual (error term from the linear models) using a LOWESS model. In the fifth case study, we test all the procedures using the firm-level stock returns (23 top companies in Fortune 500) into the same model instead of the market-level returns, revealing the impact of investor sentiment on the firm-level stock returns. Results have shown that although impact of sentiment is always observed at market-level (because the coverage of market-level sentiment is better as any news of the companies included in the index can be regarded to the market news), it is not always obvious at the firm-level (with difficulties in availability of data and limitation of method).

Acknowledgements

I would like to thank a number of people who have helped make this dissertation possible. I would like to express my gratitude and sincere thanks to Professor Khurshid Ahmad for his supervision. Without his continued support, encouragement, and expertise this thesis would not have been possible. I would also like to thank the hospitality of Trinity College Dublin, the School of Computer Science and Statistics, and the Slándáil project (EU FP7 grant #607691) for scholarship support.

I want to thank all my collogues Stephen Kelly, Xiubo Zhang, Shane Finan, and Jason Cook for their continued support, advice and friendship throughout the course of my study.

Most of all I am eternally greatly to my family, my Mam and Dad, who without their unconditional love and support I would most certainly not be here today.

A special thanks also goes to Ran, I couldn't have finished this thesis without her support.

Contents

List of Tables

List of Figures

List of Abbreviations

- **NBER:** National Bureau of Economic Research
- **VAR:** Vector Autoregression
- **GARCH:** Generalised Autoregressive Conditional Heteroskedasticity
- **LOWESS:** Locally Weighted Scatterplot Smoothing
- **i.i.d:** Independently and Identically Distributed
- **S&P 500:** Standard & Poor 500
- **DJIA:** Dow Jones Industrial Average
- **NYSE:** New York Stock Exchange
- **FTSE:** Financial Time Stock Exchange 100 Index
- **OMX(C):** Nasdaq Nordic (Copenhagen)
- **SHCOMP:** Shanghai Stock Exchange Composite
- **WTI:** West Texas Intermediate
- **WSJ:** Wall Street Journal
- **AofM:** Abreast of the Market column
- **NYT:** New York Times

- **BoW:** Bag of Words
- **GI:** General Inquirer
- **B-S:** Black-Scholes Option Pricing Model
- **ETF:** Exchange-Traded Fund
- **IMF:** International Monetary Fund
- **OECD:** Organisation for Economic Cooperation and Development
- **EMH:** Efficient Market Hypothesis
- **IPO:** Initial Public Offering
- **GDP/GNP:** Gross Domestic Production/Gross National Production
- **PMI:** Purchasing Manager Index
- **DI:** Dollar Index
- **RICI:** Rogers International Commodity Index
- **VIX:** The CBOE Volatility Index
- **API:** Application Programming Interface

Chapter 1

Introduction

1.1 Motivation and Background

1.1.1 Introduction

In fundamental finance theories, it is believed that the market is efficient and that an asset's prices fully reflect all available information. Market prices should only react to new information or changes in discount rates. However, in recent years, some researchers started to insist that investing behaviours are a social activity and it is possible that these behaviours could be affected by speculation and overreaction to events (Shiller, Fischer, & Friedman, 1984)[82]. Attitudes or fashions may influence stock prices. These trends tend to vary in different markets and appear without rational explanations. As a consequence, a boom time may be led by the irrational herding behaviour among a large number of investors and large bubbles can be created during these times (Shiller, 2000)[83]. In this sense, asset pricing is broadened by the methods based on the psychology of investors instead of the more rational way the asset prices are determined by both risks and misevaluation (Hirshleifer, 2001)[43].

The price of a financial asset has an intrinsic value[1], including the cost of discovering, handling, delivery and profit margins. If the asset is scarce or demand is strong, the intrinsic price will increase the premium; if the commodity is rich or low demand, the loss will incur during the sale. At other times, some buyers may guess what the intrinsic value of the commodity is and offer a higher price, while others may cut prices, and observing may lead to equilibrium. The transaction which triggered the price fluctuation will eventually break-even and the buyer 'discovers' a true or intrinsic price. However, irrational buyers and sellers may misinterpret each other and under- or over-estimate the commodity price.

A simple equation focuses on the above discussion: Let P_t be observed at time t, P_t^T is the intrinsic price then

$$P_t = P_t^T + \epsilon_t \tag{1.1}$$

And let ϵ_t be the fluctuation.

It is believed the value of the variable ϵ_t will be zero if the buyers and sellers are observing each other and are 'discovering' the intrinsic price. This can be denoted as the expectation value of ϵ_t:

$$< \epsilon_t >= 0 \tag{1.2}$$

At the end of the 19^{th} century, Louis Bachelier[9] observed the bond price for a period in the Paris exchange market, the fluctuation ϵ_t is a distribution according to Gaussian distribution and the expectation is zero. He was credited with being the first person to model the stochastic process which was part of his erudite and multidisciplinary PhD thesis "The Theory of Speculation" (translated from the original French).

[1]Also called a perceived or true value, which is calculated using fundamental analysis including the tangible and intangible value of an asset, which is the inherent worth of an asset. It may or may not be the same as the current market value. The opposite of intrinsic value is extrinsic value, which measures the difference between the market price of an asset and its intrinsic value.

For more than half a century, scholars have been trying to prove the existence of Efficient Market Hypothesis (EMH). The purpose of this hypothesis is to understand the instinct value of assets, commodities, stocks, bonds and so on in the market, and the impact of the buying and selling behaviours on this value.

The argument of the hypothesis is that the price has all the information that needs to be known when judging the true value of the asset. However, as mentioned earlier, that is, the market cannot control the moods and sentiment, and one can see more and more frequent prosperity and depression.

1.1.2 Markets and Calendar Effects

The invention of the market is an opportunity that helps buyers and sellers to discover the true price of goods and services: if the seller overprices the goods, they will be rejected by the buyer; if the goods are under-priced, the price is likely to raise. If the buyer and the seller cannot judge the correct price, then the market has actors, who are neither buyers nor sellers. These people rely on experience to judge the temporary malfunction of the market and these arbitrageurs quickly buy undervalued assets and sell on overvalued. Arbitrageur behaviour is an indication of the valuation strategy that will help to find the true value of the asset. There are other ways and techniques to ensure that buyers and sellers will be protected if market forces fail against the multiplicity of human behaviour with paying additional costs (also known as a premium) to purchase assets. There is also a market for insurance policies for buyers and sellers, which are buyers and sellers who need to pay the premium.

$$P_t = \alpha_1 P_{t-1} + \alpha_2 P_{t-2} + \ + \alpha_n P_{t-n} + \epsilon_t \tag{1.3}$$

where α_i is the weight of the prices in the previous i periods where (i = 1, n).

Hereafter, one can estimate the value of the weighting parameter of an asset using regression techniques as the intrinsic price of it is the weighted average of its historical prices (so long as the price does not have extreme rises and falls)(Equation 1.3). Other behaviours impact on

the prediction of stock prices using the above technique and lead to an unfair estimation of the intrinsic price of an asset such as calendar effect - the calendar effect is any market anomaly or economic effect that seems to be related to the calendar[2]. Weekends and holidays are commonly believed to have an impact as prices fall before a break and rise after it. A valuation strategy should always consider this aspect that is rooted in the calendar effects. A proxy value of 1 for the beginning and end of the week and 0 for all other days have been set to include the day-of-the-week effect, and a similar proxy has also been set for major holidays (e.g. Christmas./New year). The effect of public holidays on market share price returns are usually lower return on Fridays and higher on Mondays has been discussed widely in the literature (Taylor, 2011)[90]. Also, the impact of the loss of daylight hours and long holidays (Christmas, Easter, etc.) is also regarded as negative on price return. The calendar effect is cited as one of the key effects where the efficient market hypothesis was challenged. These holidays should not have any bearing on the price of the asset. The impact of the holidays was included in the regression equations as a dummy variable in the equation.

$$P_t = \alpha_1 P_{t-1} + \alpha_2 P_{t-2} + \ + \alpha_n P_{t-n} + \lambda CALENDAR_t + \epsilon_t \tag{1.4}$$

where $CALENDAR_t = 1$ if any calendar effect; $CALENDAR_t = 0$ otherwise. λ is the weight of these proxies for calendar effect.

Systematic indicators of market volatility and depression are described in the history of long asset value as a term: business cycles that range in length from 3 to 5 years, 7 to 11 years, 15 to 25 years, and 45 to 60 years. Recently, the date of the market's ups and downs cycles has been measured and published in the main statistical authority (NBER[3], 2010)[69], and the impact of these cycles are reflected in the valuation of the value. The proxy for the crisis is set to 1 when

[2]This effect includes the distinct behaviour of the stock market on different days of the week, different times of the month, and different times of the year (seasonal trends).

[3]The National Bureau of Economic Research (NBER) announces business cycle dates with peaks and troughs on a monthly basis.

the date is in the crisis or 0 when the date is not in crisis and is added to Equation 1.4 and shown in Equation 1.5.

$$P_t = \alpha_1 P_{t-1} + \alpha_2 P_{t-2} + \ + \alpha_n P_{t-n} + \lambda CALENDAR_t + \kappa NBER(t) + \epsilon_t \tag{1.5}$$

$NBER(t) = 1$, during or on the date of a crisis;

$NBER(t) = 0$, otherwise;

where κ is the weighting factor for the proxy for business cycles.

1.1.3 Market and Firm Sentiment

When companies list on the stock exchange, their stock prices fluctuate accordingly. It has been previously discussed that many variables for the method of discovering the intrinsic value of the stock, combined with the valuation of the intrinsic value of the asset.

Many economists (Sims, 1980; etc.)[86] have studied unforeseen events and the impact of policy changes on the economic market. The Sims method is mainly tested in a multivariate time series, where the independent variables in the equation are used to interpret the dependent variable and its linear relationship. In this method, independence and dependent variables are interchangeable, while statistical tests give significance to their correlations (Granger, 1977)[41]. There are many variables (exchange rates and interest rates, commodity prices, inflation and volatility) that will or will not affect large economies. Based on the macroeconomic approach, sentiment proxy variables can be used as independent or dependent variables for asset prices. With this vector autoregressive model, statistical tests can determine whether asset prices depend on market sentiment and vice versa. In the recent past, many economists have been trying to use a proxy to estimate market sentiment. Market sentiment usually causes asset prices to fluctuate according to their intrinsic prices. When new stocks are listed, people are uncertain about the value of new assets, so asset prices fluctuate. In addition, this uncertainty is likely to spread to other assets and other markets (Baker and Wurgler, 2006)[10].

This method is also used in the past decades for elections and voting results in the Western countries. The uncertainty of polls in such elections is very large, and the voting result is often an indication of the intention of the voters. Political parties and leaders can observe voters' sentiment by observing public opinion polls. Second, the leader also used public opinion polls to revise his speeches and performance to try to guide voters' intentions. This kind of reference index to measure public sentiment is often considered to be easier to observe and analyse than the public opinion itself.

There are macro and micro entry points for market research. The market level is a comprehensive study. The index is a particularly useful proxy for market movements. It smoothes the movement of each stock and spreads outliers. When one or two stocks fail, it is safe to use the Vector Autoregression (VAR) model to study the market; but when 100 stocks fail, this example does not hold. From the formulas provided by the Standard & Poor 500 (S&P 500) and Dow Jones Industrial Average (DJIA), it is clear that the index price is the geometric mean of the price of all participating stocks, which is an aggregation of various types of companies that are scattered. Media content reflects the company's fundamentals and is quickly incorporated into stock prices by investors (Tetlock, 2007; Tetlock, SaarTsechansky, & Macskassy, 2008).

1.1.4 Proxy of Sentiment

Over the last two decades, researchers investigating complex political, economic and financial systems have discovered that sentiment is an important measuring tool. In this thesis, it is focused on financial systems. Investor sentiment is the general prevailing attitude of investors as to anticipated price movements in the market. However, it is not directly quantifiable. Most researchers have carefully sourced and elaborately constructed proxies representing investor sentiment and eventually used these as measures for analysing variations in stock market prices.

Previous studies computed the so-called investor sentiment by investigating the number of messages in a message board or the frequency count of pre-selected words in newspaper texts. The market movement is proxied using an aggregate index which is a measure of the value of a

section of a stock market; the assumption here is that the aggregate index will reflect a consensus about the state of the market. Recently, many studies have focused on investor sentiment in media sources (including traditional and social media) and its effect on the changes in the indices. Negative sentiment has been found to significantly influence stock market movements.

However, despite the volume of behavioural finance studies over the last few decades, there are some questions (problems) that remain unsolved. Firstly, most researchers have worked on the US markets and have rarely been concerned about the role that investor sentiment plays in other markets. Secondly, although researchers have successfully detected the impact of investor sentiment on the stock market, most of these studies are historical with low-frequency data. Thirdly, in textual sentiment analysis, English is the most used language, and studies conducted in other languages are uncommon. This research attempts to address each of these three largely unexplored problems.

The potential contributions of this research are 1) universality of investor sentiment and its impact that are tested on the US market using *Wall Street Journal*'s "Abreast of the Markets" column and further extended to Danish and Chinese markets using news reports from newspapers and news agencies and to confirm whether investor sentiment affects price returns on a regional, macro and data diversity basis; 2) analysis of the sentiment and its impact is extended to carefully selected top US companies; 3) analysis of residuals of multivariate regression are conducted to observe and confirm the nonlinear relationship between investor sentiment (extracted from stock indices and top companies) and return residuals by locally weighted regression and 4) the development of a system for financial investors/researchers to get a deep understanding of the relationship between sentiment proxies and financial returns.

1.2 Research Questions

A market can be viewed at different levels of description: one could look at aggregating the indices carrying all or part of the market (SP500, DJIA); or a firm grain view will be to look

at individual companies or for multinationals, like IBM, Shell, and BP which are in different markets, one can look at the same company in different markets.

Markets as a whole can be viewed in terms of scale, New York market vs Copenhagen market; or an established vs emerging markets, again New York vs Shanghai; or Shanghai vs Copenhagen. The question arises whether sentiment is of significance at the company level, or at the market level.

Sentiment is supposed to be included in texts of a newspaper and there is some evidence for this at the market level. However, the texts are usually a commentary or opinion. But there are different types of texts - the so-called news cline.

The principal question this thesis attempts to answer is:

- Can a method be developed to systematically detect the existence and impact of investor sentiment at different levels of market (aggregate and firm levels)?

1.3 Key Conclusions

The focus of this thesis is to create a computational solution for calculating the proxy of the information contained in the news content and to evaluate the influence of this proxy on changes in the financial market (including the index level and the firm level). In past literature, methods of text and sentiment analysis have been used in financial literature to generate such proxies from news. Research and systems developed to accomplish this task first selected a financial market for analysis, with stocks and indices being the most widely studied. This thesis first reproduces the results of Tetlock (2007) sentiment analysis of DJIA (using the Abreast of the Market (AofM) column in Wall Street Journal from 1984-1999 as a text source). This research was in-depth and the data was extended to 2015 (AofM column was redesigned as a weekly column in 2007. This thesis replaces the post-2007 period with the Wall Street Journal and the New York Times' commercial and economic news.) to verify the dependency of using such method. From a different perspective, a similar study was also conducted on a small European economy (Denmark) and an emerging economy (China). The business and economic news was

obtained from local newspapers in Denmark and China from 2000 to 2015 (an English translation of the local language) to extract sentiment and conduct sentiment analysis. This multi-market study has advanced the understanding of the effects of sentiment in different markets. Few studies in the previous literature have conducted time series analysis of company-specific text sentiments, and few have studied the time-varying model of its role in the stock market. This is because it is difficult to obtain continuous sentiments for individual companies. This thesis has created a series of sentiment proxies time series with daily frequency for 23 large companies for 15 consecutive years.

The prototype developed in this thesis includes a series of tasks, including data acquisition and collection, data processing and aggregation, content analysis, time series modelling, model diagnosis and hypothesis testing, and result visualisation. The content analysis method and framework chosen for implementation allows the collection of any text corpora and financial time series data and its introduction to estimating the impact of news on financial returns.

The contribution of this thesis can be summarised as follows: First, this thesis has formed a process to analyse sentiment impact through well-designed methods and procedures. Second, this thesis collected a data set to see the aggregate index - sentiment on all. Third, this thesis provides a collection of search strategies - recommendations for keywords and company search methods. Fourth, a method has been designed for classifying sentiment behaviours. Fifth, the results have been categorised arranging different indices and companies according to the size of the capital (confirmed that the size of the index does not matter, the sentiment impact remains stable; but there will be deviations at the company level). Sixth, this thesis studies the contribution of the preparation of the corpus (the Bag of Words (BoW) model cannot fully cope with the company-level sentiment analysis, company news has overlapping within the industry, and the current technology is very hard to accurately separate them).

1.4 Publications

- Zhao, Z., & Ahmad, K. (2015a). Qualitative and Quantitative Sentiment Proxies: Interaction between Markets. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 466-474). Springer International Publishing.[103]

- Zhao, Z., & Ahmad, K. (2015b). A Computational Account of Investor Behaviour in Chinese and US Market. Int. J. Econ. Behav. Organ, 3(6), 78-84.[104]

- Cook, J. A., Zhao, Z., & Ahmad, K. (2016). Stylised Facts of Linguistic Corpora: Exploring the Lexical Properties of Affect in News. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 494-502). Springer International Publishing.[25]

- Zhao, Z., Kelly, S., & Ahmad, K. (2017). Finding Sentiment in Noise: Non-linear Relationships Between Sentiment and Financial Markets. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 580-591). Springer, Cham.[105]

1.5 Thesis Structure

A review and discussion of the literature using text analysis methods with particular attention given to content analysis is presented first. Time series models used in this thesis are also described with their application in econometric models. Finally, studies and systems combining methods of text analysis and financial analysis and prediction are then discussed (Chapter 2).

A detailed description of all methodologies, an analysing approach and its implementation is described (Chapter 3). The main components of the approach, the text analysis component and the statistical modelling component and the necessary data harvesting and preprocessing functionality are described.

The computational approach is then evaluated by conducting five case studies with text and time series data for market- and firm-level price returns(Chapter 4). By employing parametric

statistical methods such as vector autoregression, rolling window regression, locally weighted regression and hypothesis testing, the explanatory information of the sentiment variable computed by the approach from the text corpus for the different financial assets is assessed. How this sentiment variable impacts financial returns is investigated in different indices, equities, commodity markets and firm level stocks. The influence that different text types and text corpora have on the content analysis method are also evaluated by looking at the role different text sources and news article types play on the computation of the sentiment time series variable. An investigation into the changing influence of sentiment over time and in different market conditions is examined for each financial market in the case studies. Lastly, a semi-parametric approach[4] is conducted to observe the relationship between negative sentiment and return residuals at both market and firm levels.

The final chapter of the thesis summarises the work presented giving a brief discussion and commentary, concluding with a discussion of future work (Chapter 5).

[4]An approach consists with two stages - a parametric approach to generate return residuals followed by a non-parametric to test the relationship between residuals and sentiments.

Chapter 2

Related Work

2.1 Quantitative Awareness

2.1.1 Behaviour of Matter

The history of financial mathematics can be traced back to the 1900's French mathematician Louis Bachelier's defended his doctoral thesis "The Theory of Speculation". In the text, he first used Brownian motion to describe the change in the stock price. He thought that, because there is both a buy and a sell in the capital market, the buyer was bullish and the seller was bearish, and that the price fluctuation was Brown's movement and the distribution was normal. The beginning of modern finance followed two major revolutions, the first in 1952. That year, 25-year-old Markowitz published his doctoral thesis, proposing the theory of mean-variance of portfolio selection[1]. Its meaning is to guide the idea that people are looking for "best" stocks to understand the quantification and balance of risks and profits. Given the risk level maximises the expected return (or given the level of income minimises the risk) is the main idea of the above mean-variance theory. Later, Sharpe[81] and Lintner[57] further expanded the work of Markowitz, proposed the Capital Asset Pricing Model (CAPM), followed by Miller's presentation[63] of the

[1]Also known as the Modern Portfolio Theory, is a process to weight risk (or variance) and maximise return by diversification in investment.

company's financial theory (MM theory) that sparked the first "Wall Street Revolution" in the 1970s, which is the beginning of financial mathematics. Markowitz and Sharpe also won the 1990 Nobel Prize in Economics for their pioneering contributions to financial mathematics.

2.1.2 Dealing with Returns

In the 1960s, based on Markowitz's mean-variance model, Sharpe (1964)[81], Linter (1965)[57], and Mossin (1966)[64] independently proposed the well-known Capital Assets Pricing Model (CAPM), that is, in the capital market equilibrium, any securities or portfolio of securities and the risk of the following linear relationship:

$$r - r_f = \beta(r_M - r_f) \tag{2.1}$$

where r is the securities or the expected rate of return of the portfolio, r_f is the risk-free rate of return, r_M is the expected rate of return for the market portfolio, and β is the systemic risk measure for the portfolio of securities. CAPM is considered to be the backbone of the modern financial market price theory, and is widely used in the determination of portfolio performance, securities valuation and capital cost calculation. Sharp won the 1990 Nobel Prize in Economics for his pioneering achievements in this field.

The CAPM model, when considering the factors that determine the rate of return on assets, only analyses one factor, the impact of the market mix, and the assumptions required by the CAPM model are too strong. In view of this, Roll and Ross (1980)[73] proposed an Arbitrage Pricing Theory (APT) from the development of the capital asset pricing model. APT is based on a monopolistic law, and its theoretical point is that the yield of securities is linearly related to a set of factors that affect it, ie

$$R_i = a + b_a F_1 + b_a F_2 + ... + b_q F_j + g_i \tag{2.2}$$

where R_i is the i^{th} stock F_j is the j^{th} factor that affects the yield of the securities, b_q is the sensitivity of the yield i of the securities i to the element j, and g_i is the random error term.

2.1.3 Random Motion of Financial Market and its Volatility

The random error term has been properly investigated by researchers starting with its unpredictable nature using the famous physical theory - Brownian Motion. In 1973, Black and Scholes[17] introduced the option pricing formula using mathematical methods. Furthermore, Morton[61] developed and deepened the formula. The option pricing formula brings convenience to financial traders and bankers in the transaction of derivative financial assets, which promotes the development of option transactions. Option trading quickly becomes the main content of world financial markets and becomes the second "Wall Street Revolution" in the 1970s.

Two "revolutions" to avoid the general economic equilibrium of the theoretical framework of finance, formed a new interdisciplinary area. The Markowitz-Sharpe theory and the Black-Scholes (B-S) formula together constitute the new subject of vigorous development - the main content of financial mathematics, but also the new theoretical study of new derivative securities - the theoretical basis of financial engineering. The pioneers of the revolution won the Nobel Prize in Economics in 1990 and 1997 respectively. American economist Robert Engle and British economist Clive Granger[34] on the time series theory in economic and financial research achieved remarkable results further in 2003 they won the Nobel Prize in Economics, and it is worth noting that this is the third time that mathematical financial research has won the Nobel Prize in Economics. This emerging discipline, Financial Mathematics, has become a magnificent work of the international financial community.

Since the option pricing model gives the quantitative relationship between the option price and the five basic parameters (S_t, K, r, $T-t$ and σ), one can substitute the first four basic parameters and the actual market price of the option as the known quantity in the option pricing model. One can then solve the only unknowns parameter σ and its value is the implied volatility. Therefore, the implied volatility can be understood as the expectation of the actual volatility of

the market. The option pricing model requires the actual volatility of the price of the underlying asset over the life of the option.

$$C(S_t, t) = N(d_1)S_t - N(d_2)Ke^{-r(T-t)} \quad (6)$$
$$d_1 = 1/\sigma\sqrt{T-t}[ln(S_t/K) + (r + \sigma^2/2)(T-t)]$$
$$d_2 = d_1 - \sigma\sqrt{T-t}$$

$$P(S_t, t) = Ke^{-r(T-t)} - S_t + C(S_t, t)$$
$$= N(-d_2)Ke^{-r(T-t)} - N(-d_1)S_t \quad (7)$$

The above are formulas for calculating call and put option price, respectively, given by the B-S formula where

$T - t$ represents the maturity to the present time interval;

S_t denotes the price at time t (at present);

K is the strike price;

r stands for the risk-free interest rates;

σ is the volatility of return.

It is an unknown quantity relative to the current period. Therefore, it needs to be replaced by forecasting volatility. Historical volatility estimation can be simply used as forecasting volatility. However, it is necessary to use quantitative analysis and qualitative analysis combined with the historical volatility as the initial forecast value, continue to adjust the correction, to determine the volatility based on quantitative data and the new actual price information.

Volatility is the fluctuation of the price of a financial asset and is a measure of the uncertainty of the return on assets and is used to reflect the level of risk in a financial asset. The higher the volatility, the more volatile the price of financial assets, the greater the uncertainty of a return on assets; the lower the volatility, the more stable fluctuations in the prices of financial assets, the greater the certainty of a return on assets. In the sense of economics, the main reason for such volatility existing comes from the following three aspects:

1. the impact of macroeconomic factors on an industry sector, namely system risk;

2. the impact of a particular event on a single business, called non-system risk;

3. the effect on changes of psychological status or expectation of investors in the stock price.

The existence of volatility has been proved by a number of studies (Taylor, 2011; Bartram, Brown & Stulz, 2012)[90][13]. Volatility describes the standard deviation of returns over some period of time (Taylor, 2011)[90]. There are two types of volatility, one is backwards-looking, and the other is forward-looking. The former is calculated using the historical data of volatility. The latter is based on the current option price, using the B-S option pricing model to derive the volatility. The former is a history of price fluctuations that have taken place and from which is calculated a volatility. The latter is a prediction of the volatility of a price in the future and may not be accurate.

2.1.4 Calendar Effects

The average stock return varies significantly over time. This difference depends very much on the day of the week, on the day of the month, in the month of the year and/or on holidays. The difference is even related to the location of the Sun, Earth and the Moon. In the past decade, many scholars have studied such calendar effects, including the Monday effect (Rubinstein, 2001; Sullivan, Timmermann and White, 2001; Schwert, 2003)[75][88][79], the Day-of-the-Week effect (Fields, 1931; Cross, 1973;French, 1980)[37][26][38], the Month-of-the-Year effect (Rozeff and Kinney, 1976; Praetz,1973;Officer, 1975; Gultekin and Gultekin, 1983)[74][71] and the Holidays effect (Ariel, 1990; Lakonishok and Smidt, 1988; Ziemba, 1991; Kim and Park, 1994)[8][54][106][50].

The finance literature (Damodaran, 1989)[30] has discussed the calendar effects, but mainly in regard to data mining (Sullivan et al., 2001)[88]. Although studies showed that the 'Monday effect' is statistically significant. However, it 'disappeared' after 1980's (Alt et al., 2011)[3].

Calendars affect the new market relatively. For instance, there is evidence that the day of the week and seasonal holidays in East Asia (except January) have an impact (Seif et al., 2017)[80]; China's New Year has a major impact on the Chinese stock market (Yuan and Gupta, 2014)[101].

Holiday influence (including the regional holidays in Europe and the United States) is often mentioned in recent years and is thought to have an impact on the stock market. It seems that this effect also occurs in other regions of the world (Lu et al., 2016)[60].

These calendar implications have important implications for the economic/financial markets in China and East Asia. Another common phenomenon is that the market will fluctuate sharply before key policies and regulatory or statistical data (directly or indirectly related to prices) concerning different economic aspects are announced (Andersen and Bollerslev, 1998, Chan et al., 2017, Birz and Dutta, 2016)[4][22][15]. The buyers or sellers will have more difficulties on the announced policy and its impact on the price of the asset, which makes the valuation more difficult.

2.2 Information Revolution

2.2.1 More Information, More Problems

Prior to the advent of modern presswork in the 14^{th} century, the reproductive cycle of handmade papers restricted the accumulation and transportation of knowledge. The cost of reproduction of information decreased by about three hundred times when the printing press was introduced and then further reduced when online press became popular since the late 1990s. Data mining mechanisms expressed the option to gather substantial information in the "Big-Data" age. The question that this thesis explores is where is the sentiment and what is the evidence of its existence.

"The original revolution in information technology came not with the microchip, but with the printing press." (Silver, 2012)[85] Scribes used to produce a copy of a book at a time. The cost for reproducing a single copy was about one florin per five pages, so a normal book (with around 500 pages) people are reading today would cost around $20,000. Due to unavoidable human mistakes, many of them carry transcription errors from one copy to another. The cost of pursuing knowledge was exorbitant with uncertainties. This has been changed since a system of

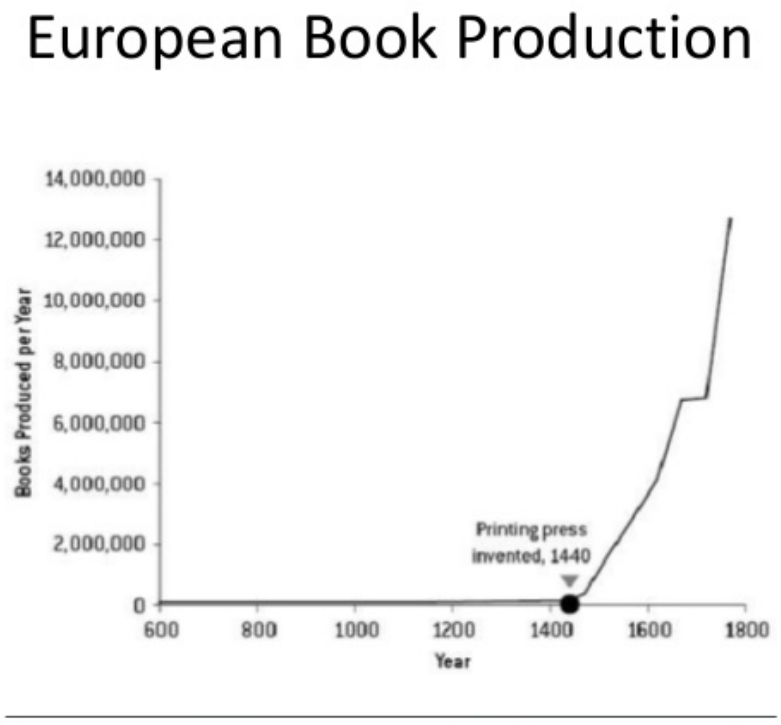


Figure 2.1: **European Book Production** (Silver, 2012)[85].

printing and typography that uses movable components to reproduce the elements of a document on the medium of paper was first invented by Bi Sheng in China around 1040 (Tsuen-Hsuin and Needham, 1985)[95] and then the metal movable-type printing press was introduced by Johannes Gutenberg in Europe around 1450 (Lehmann-Haupt, 1966))[55]. As a consequence, the cost of book transcription was reduced to less than half a percent as of what it was (instead of $20,000 today it's around $70) (See Figure 2.1). This technology hence dramatically enhanced the ability to carry or share data and guided the modern digital data evolution.

2.2.2 "Big Data": Advantage and Disadvantage

The emerging discipline of behavioural finance is not only used to explain the psychological and behaviour of micro-individuals, but also the study of the financial market where it has a very wide range of applications. For example, many hedge funds adopt "inertial strategy", which

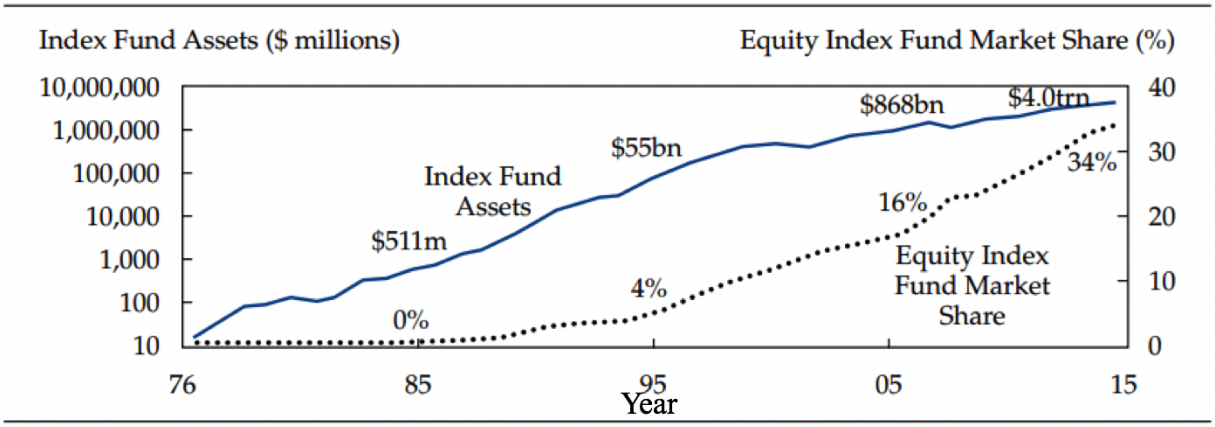


Figure 2.2: **The Growth of the Index Fund.**(Bogle, 2016)[18]

takes advantage of the fact that investors do not reflect enough market information to make profits.

Artificial intelligence has begun to take the place of artificial labour in the financial industry. The BlackRock, the world's largest asset manager, has recently laid off seven pro-active fund managers and replaced them with a quantitative investment strategy. Larry Fink, the founder of BlackRock, said in an interview (Toonkel, 2015)[94] that future investments will rely more on big data, artificial intelligence, quantification and factors and models in traditional investment strategies. Over the past decade, the growth of index funds and Exchange Traded Funds (ETFs) (Bogle, 2016)[18] has far outstripped that of actively managed funds. From this point of view, raised fund managers are facing the fact that they are being gradually replaced. Perhaps before the advent of the true artificial intelligence era, many financial practitioners will have been replaced by quantitative or indexed funds and ETFs (See Figure 2.2).

There are two very obvious advantages to using a machine or artificial intelligence. First, it is less likely to make some very simple mistakes. Second, it should not be affected by the human emotions that can affect the investment decision-making process. If artificial intelligence is trained in the field of behavioural finance, the most important application direction is to avoid the mistakes that human investors often make such as overconfidence, mental account, aversion loss, herding effect and so on. More importantly, machines can be trained to take advantage of the weaknesses of human nature to profit for themselves.

Will the application of artificial intelligence bring the market risk? Thirty years ago, in 1987, the largest financial calamity in history - "Black Friday" - took place in the United States. One of the important reasons for the stock market crash was the heavily used programmatic trading on Wall Street: the "portfolio insurance" strategy. When the price spread[2] expanded, the spot arbitrage[3] mechanism was triggered automatically to buy the futures and sell the spot goods[4]. It also shows the consistency of programmatic transactions, so that the collapse became more violent and spread rapidly. If at a particular point in time, all artificial intelligence make the same judgement that all stocks are sold during the centralised time, it will not lead to a stock market crash that should not have occurred.

2.2.3 The Prediction

In the finance community there exist different theories and assumptions about market sentiment, investor behaviour, and its role in financial markets. Traditional beliefs held in finance of an efficient market with rational participants are not as forthwith as initially theorised. The strict idea of an efficient market, that all information is incorporated into price at all times, would lead investors to believe that news and exogenous information would have no impact on the markets, having already absorbed any and all possible innovations. The limits of the Efficient Market Hypothesis have since been accepted since the idea that a market contains all relevant information seemed unfeasible, with evidence against the theory supported by irrational market behaviour (Black, 1986)[16]. The theory was revised to include additional forms of market efficiency namely strong, semi-strong and weak forms of efficiency (Fama, 1970)[35]. The limits and doubts of the efficient market hypothesis were also posed by Kahneman and Tversky (1986)[96] who highlighted the importance of the behavioural aspects of financial markets and the role investor sentiment plays, specifically investors ability to overreact to events.

[2]The different between the price and the market value.

[3]An opportunity to profit from the mispricing between the spot and future price of the underlying assets.

[4]Spot price is the current market value of an asset; future price is the price at which the asset can be bought or sold in the future under a contract/agreement.

Studies in the finance community have attempted to measure and create a proxy for sentiment contained in financial markets in an attempt to detect the onset of irrational behaviour or unfounded speculation. The issue becomes how to measure sentiment and incorporate it into a model that can reliably explain changes in the price of an asset. A number of different proxies for investor sentiment have been proposed and studied in financial literature including surveys, trading volume, order flow, dividend premiums, among others (Baker and Wurgler, 2006)[10]. More recently, content and text analysis methods have been used to extract sentiment from news and text published online in an attempt to summarise the mood of investors towards changing markets (Tetlock, 2007; Garcia, 2013; and Kelly and Ahmad, 2015)[92][40][48].

Using content analysis methods to analyse text means a large volume of qualitative information can now be incorporated into a quantitative model. Sentiment analysis applied to news means the tone of the text can be summarised and this may influence a wide readership invoking a response that might be reflected in asset prices. Previous studies that have relied on sentiment analysis to summarise sentiment from text have found that negative sentiment can reliably predict changes in market indices (Schumaker and Chen, 2009; Tetlock, 2007; Garcia, 2013)[76][92][40], stock prices (Jegadeesh and Titman, 1993; Tetlock, Saar-Tsechansky, and Macskassy, 2008)[44][93], and commodity markets (Feuerverger, Yu, Khatri, et al., 2012; Kelly, 2016)[36][47]. These studies have identified a linear relationship between their measure of news sentiment and financial returns for a range of different assets. The work presented here follows from these studies to identify this linear relationship between sentiment and financial assets in the Chinese and Dutch equity markets. One shortfall of these studies is the lack of identifying potential non-linear effects. In this thesis, the non-linear effects have been inspected by examining the residual series produced from incorporating news sentiment and returns together into a single explanatory regression-based model. Following from Tetlock (2007)[92] a consistent non-linear relationship at the index level and selective relationship at the firm level between the residual and sentiment have been identified. It is found that this relationship is across a number of different financial markets showing sentiment to have had an influence on returns in certain periods.

2.2.4 Efficient Market Hypothesis and Behavioural Finance

In 1970, Fama[35] came up with the efficient markets hypothesis (EMH), which defines the efficient market as such: If the price in the stock market completely reflects all the available information then such a market is said to be an efficient market.

There are three forms of valid capital market hypothesis: Weakly efficient markets, which completely contain information about past prices, make technical analysis useless. Semi-strong, efficient market with all publicly available information, making most financial analysis useless. A strong, efficient market that contains all the public and inside information and no one in the market can take advantage of early information anymore.

There are two signs to judge whether the stock market has extrinsic efficiency: First, whether the price can be freely changed according to relevant information; Second, whether the relevant information of the securities can be fully disclosed and evenly distributed to each investor at the same time.

The conditions for becoming an efficient market are:

(1) Investors all use available information to seek higher profit;

(2) The stock market's response to new market information is rapid and accurate, the price of securities can fully reflect all the information;

(3) Market competition makes the price of securities transition from old equilibrium to new equilibrium, and the price changes corresponding to new information are independent or random.

The theory includes the following points:

First, everyone in the market is rational. Every company represented by each stock in the financial market is under the rigorous surveillance of these rational people. They conduct a basic analysis of the company's future profitability to evaluate a company's share price, convert future value to the present value on the day, and carefully weigh trade-offs between risk and return.

Second, the price of the stock reflects the balance between the supply and the demand of these rational persons. The person who wants to buy is exactly equal to the person who wants to sell. That is, the person who believes the stock price is overvalued equals the person who believes the

stock price is undervalued. When the two are not equal, there is a possibility of arbitrage, they immediately change their stock price to the same price by buying or selling the stock.

Thirdly, the price of the stock also fully reflects all the available information of the asset, namely, "information is valid." When the information changes, the price of the stock will certainly change. When good or bad news emerges, the price of the stock starts to shift, and when the good/bad news was well known, the price of the stock had risen or dropped to the proper price.

Behavioural finance is an interdisciplinary discipline of finance, psychology, behaviour and sociology, trying to reveal the irrational behaviour and decision-making rules of financial markets. Behavioural finance theory holds that the market price of securities is not only determined by the intrinsic value of securities, but also largely influenced by the behaviour of investors, that is, the psychology and behaviour of investors have a significant influence on the price decision of securities markets and its changes. It is a theory corresponding to the efficient market hypothesis. The main contents can be divided into arbitrage restrictions and psychology.

The traditional efficient market hypothesis holds that the price in the financial market contains all the information, and at any time the price of securities can be regarded as the best estimate of the investment value. According to the theory of behavioural finance, there are two assumptions about the behaviour of investors in the EMH: First, there is no deviation in the behaviour patterns investors take when maximising the value of the portfolios they own. Second, investors always aim to maximise their own interests.

Behavioural finance believes that the efficient market hypothesis itself does not guarantee that these two premises must be established. On the contrary, behavioural finance has questioned the correctness and rationality of these two hypothetical premises based on an analysis of the actual situation and believes that investors often violate these two hypothetical premises because of psychological factors. Traditional theory failed to take into account the subjective errors caused by the psychological factors of fund managers and investment mistakes is the obvious flaw, the psychological factors should be the choice of funds to invest and selecting fund managers is a very important consideration.

In summary, the efficient market hypothesis is the foundation of classical economics as a hypothesis that reflects the ideal state that economists and financial economists dream of. Although there are all kinds of visions in reality that challenge the Efficient Market Hypothesis, they reflect the deviation of the real state from the ideal state. The challenge does not fundamentally deny the efficient market hypothesis. Behavioural finance seems to help us to explain the visions of efficient market assumptions, but only partially explain the paradox of efficient market assumptions from the actual psychological activity of people, with too many uncertain variables.

2.3 Note on Behavioural Finance

In fundamental finance theories, it is believed that the market is efficient and that an asset's prices fully reflect all available information. Market prices should only react to new information or changes in discount rates. However, in recent years, some started to insist that investing behaviours is a social activity and it is possible that these behaviours could be affected by speculation and overreaction to events (Shiller, Fischer and Friedman, 1984)[82]. Attitudes or fashions may influence stock prices. These trends tend to vary in different markets and appear without rational explanations. As a consequence, a boom time may be led by irrational herding behaviour among a large number of investors and large bubbles can be created during these times (Shiller, 2000)[83]. In this sense, asset pricing is broadened by the methods based on the psychology of investors instead of the true rational way the asset prices are determined by both risks and misevaluation (Hirshleifer, 2001)[43].

An investigation into how a cross-section of stock returns have been affected by investor sentiment was carried out in (Baker and Wurgler, 2006)[10] using different proxies for market sentiment. It was found that returns, including small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks, are proportional to beginning-of-period proxies for sentiment. Their investigation demonstrated the value of adding a proxy for sentiment in explaining and accounting for temporary variation in prices.

The impact of country-specific news on the closed-end country fund prices to asset value was tested (Klibanoff, Lamont and Wizman, 1998)[51]. The results indicate that news events evoke quick reactions from investors. Prices are impacted much more by news appearing on the front page of the New York Times than stories appearing throughout the paper. This demonstrates that investor and market participants attention is fixated and influenced by news while their reactions have not fully been incorporated into price.

Several studies in recent times have relied on a quantitative measure of sentiment that is extracted from text and news using text analysis techniques. Many of these studies relied on using content analysis programs such as the General Inquirer (Stone, et al., 1966)[87] to track the frequency of sentiment laden words as they appear in a collection of news. The collection of news articles, or corpus, was constructed to have a unifying theme such as financial news or oil news. Other text types were used such as 10-Ks (Loughran and McDonald, 2011)[58], online messages boards, and more recently social media (Bollen, Mao and Pepe, 2011)[19]. In many of these studies sentiment extracted from relevant news and text was seen to have predictive power for financial returns and determining price trends.

In many of the main studies from the domain of finance, negative sentiment is seen to have the strongest influence on market prices (Tetlock, 2007)[92]. Tetlock defines a pessimism parameter based on the presence of negative news and finds it predicts a 4 basis point change (0.04%) in the Dow Jones Industrial Average. Following this study, the effect of investor sentiment on asset prices is examined in Garcia (2013)[40] for the period between 1905 and 2005. Garcia looks at DJIA returns and controls for some factors that may be a source of sentiment in a time-series regression model. It has been found that stock returns are better explained when sentiment is incorporated into the model. A one standard deviation news sentiment impact causes 12 basis points decrease in DJIA returns during recessionary periods.

A more recent study that looked at national newspaper strikes in several countries, modelled using regression based analysis, showed that the dissemination of information among investors improved when media content was incorporated into stock prices (Peress, 2014)[70].

News effects are not consistent with the Efficient Market Hypothesis (EMH) in the long-term. Evidence for this has been seen in the US, the UK and Dutch markets incorporating national news, and with three Dutch banks during the financial crisis (Schumaker, et al., 2012)[77]. Greater transparency may help in order to limit panic caused by news announcements in the global financial market.

Many studies have relied on linear regression based models to aggregate and model sentiment with financial returns. One can take this approach and use Vector Autoregression (VAR) to estimate the linear relationship between the sentiment measure and returns in several different financial markets. Taking this approach allows a model that is easy to estimate and interpret, that gives an indication of the statistical significance of the impact of sentiment on returns. The work was extended from these studies and in particular in Tetlock (2007)[92] to examine the potential non-linear relationship between sentiment and returns. The variance that is not accounted for by this linear model may be contained in the residual of the model such as non-linear effects. By examining this residual, one can determine further relationships with the sentiment variable.

Price return of an asset at time, (r_t), has a number of interesting properties, including the auto-regressive nature of dependence of prices at one moment in time with an n past price returns; the weighted average of lagged returns, denoted symbolically as a lagged operator $L_n r_t$, and denotes the symbolic weight:

$$\alpha L_n r_t = \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + ... + \alpha_n r_{t-n} + \epsilon_t \tag{2.3}$$

can help in the estimation of forecast return (r_t). The error in the estimation is denoted as ϵ_t; the error is generally assumed to be normally distributed in time and has a zero mean and unit standard deviation $(NID(0,1))$. This model is based on a number of assumptions one of which relates to the closed nature of the model-forecast return which depends only upon returns. This notion of using other correlated variables has been advocated by leading econometricians (e.g. Sims (1980)[86]) for macroeconomic analysis in general. Sims asserts that the other variables may be used in detecting the influence of the public's change of taste, a subjective argument

at its best one may argue, and this change introduces what he calls 'disequilibrium economics': This disequilibrium idea is one of the fore-runners of modern sentiment analysis. More of this shortly. It has been argued that changes of taste/fashion are indicated by initial public offerings (IPO) on the market, especially the short-term and large scale fluctuations in price returns of an existing asset due to unrealistic demands for the IPO firm (cf. Dotcom boom/bust in 2000); a number of IPOs has been cited as a proxy for investor sentiment, and following Sims have been added to a model of asset returns to forecast future returns of an asset (Baker and Wurgler, 2006)[10]. Business cycle fluctuations may also be due to calendar effects - asset trading is pursued mainly on weekdays and during daylight hours (the average volume of assets traded at other times is generally rising but still is small in comparison to the normal Monday-to-Friday, 0900-1700 turnovers), trading is impacted by public holidays, especially before and after long holidays - January trading volumes are lower due to the post end-of-Gregorian calendar a month previously. These fluctuations can be added into Equation 2.3 as dummy exogenous variables - the inclusion of exogenous variables in forecasting returns has a rich history of arguments of pros and cons of including these variables and have been discussed since the introduction of these variables about 50 years ago (Balestra and Nerlove, 1966)[11].

Sentiment analysis focuses on how the basic market dictum, that price (return) of an asset contains all the necessary information relevant to its potential for buying/selling, breaks down and one has to rely on other sources of data. For example, a new invention related to (an individual or) a firm may lead to a mispricing of the asset; during wars and periods of economic boom and bust there are many instances of over- and under-pricing, or in anticipation of a central bank announcement or before/after a summit meeting asset prices change irrespective of other critical economic performance data (Engle, 2003)[33]. These three instances, that of high-frequency, medium frequency, and low frequency respectively, show that one has to rely on other sources of data. It has been suggested that one has to assume that the key stylised facts, especially first and second moments of return will show a time-dependence; one theoretically well-founded model for dealing with this heteroskedastic behaviour is the generalised auto-regressive model that will relate the asset price returns to volatility measures like mean and standard deviation.

Alternatively, other empirical volatility measures can be added on to Equation 2.3 for a well-grounded estimate of the forecast return: lagged squared returns ($L_n r_t^2$), lagged (detrended) volume ($\rho L_n Vlm_t$), or other option-traded volatility indices are used as like VIX_t are used in an empirical fashion and added to Equation 2.3 above. Other historic indicators of volatility, especially dates of market crashes, include major non-forecasted upturns and downturns in financial and/or commodity markets. The vocabulary of papers in econometrics and finance repeatedly used the terms news arrivals, inventions, innovations, news flow that can be broadly associated with a larger than expected value of the residual ϵ_t. This, in turn, implies that at these instances in time the forecasted value of return (equation 2.3) was not accurate. The addition of the exogenous variables of different hues, including in some cases lagged values of other variables, is ostensibly designed to minimise the value of the residual terms and thereby improve the accuracy of the forecast return.

The 1990s innovation, based on the earlier work of Philip Stone and his predecessor Harold Laswell in political science in the 1940s (Stone et al., 1966)[87], led to the harvesting of sentiment from opinion pieces published in upmarket and business oriented (daily) newspapers like The Wall Street Journal (Tetlock, 2007; Garcia, 2013)[92][40] and sometimes the New York Times. The phrases like news arrivals etc., were closer to their everyday meaning. The choice of opinion pieces, comprising often unattributed remarks on the behaviour of an individual or a firm, is an interesting one as these near-gossip column factoids were expected to influence investor sentiment. Furthermore, until the 1990s sentiment harvesting was conducted by a bag-of-words (BoW) model that uses a thesaurus of emotions created by Stone et al.[87] that is essentially a word list classified (simultaneously) according to many emotion categories - sentiment being a sub-category of emotion and in Stone and Laswell's world sentiment was an evaluative word indicating positive or negative evaluation; and there are other emotional dimensions including strength and activity.

The other evolution on harvesting investor sentiment has come from scholars who have opted for a lower frequency harvesting of sentiment from documents generated by a firm to report on its own health and well-being. Scholars have used newer specialised thesauri (Ahmad, Daly and

Liston, 2011; Ahmad et al., 2014; Zhao and Ahmad, 2015)[1] [2] [104] on sentiment harvesting from news reports and opinion pieces, whilst others have used the low-frequency documents (Form 10-K in the USA filed every six months by individual firms) together with an up-to-date thesaurus (Da, Engelberg and Gao, 2014)[29]. The impact of negative sentiment on return prices is of the order of 5-10 basis points for stocks and indices, and an even greater impact of sentiment has been reported on commodity prices (Kelly, 2016; Murphy, Kelly and Ahmad, 2015)[47][66].

The information is tabulated (see in Table 2.1) with the regression schemes used by the various authors discussed above and this show that (i) following Sims (1980)[86] these authors have added more and more correlated variables; (ii) following Balestra and Nerlove (1966)[11] have added interesting exogenous variables as new vectors to an existing auto-regression scheme; and (iii) following Engle (2003)[33] either second moment of the price return, option price index volatility, or a heteroskedastic measure of volatility has been used.

Table 2.1: **Inclusion of sentiment and volatility:** E denotes exogenous variables, S denotes sentiment variables, V_1 denotes volatility-covariant measure and V_2 denotes implied volatility measure; DW, J, 1987, NBER, EP and BC shorts for the Day of the Week dummy, the January dummy, the 1987 crash index dummy, the NBER recession dates dummy, the Economic Policy News proxy and Business Conditions Index respectively.

Return	Regression scheme	Eq# (2.4)	Reference	Exogenous variable
	$r_t = \alpha L_5 r_t + \epsilon_t$	(2.4a)		
E	$r_t = \alpha L_5 r_t + \lambda_1 Exog_t + \epsilon_t^i$	(2.4b)	[11]	NA
S+E	$r_t = \alpha L_5 r_t + \beta L_5 s_t + {}_1 Exog_t + \epsilon_t^{ii}$	(2.4c)	[92][93]	DW+J+1987
S+E+V_1	$r_t = \alpha L_5 r_t + \xi L_5 r_t^2 + L_5 s_t$ $+{}_1 Exog_t + \epsilon_t^{iii}$	(2.4d)	[40]	DW+NBER
S+E+V_2	$r_t = \alpha L_5 r_t + \xi VIX_t + s_t$ $+{}_1 Exog_t + \epsilon_t^{iv}$	(2.4e)	[29][98]	EP+BC

2.4 Fusion of Sentiment and Return

2.4.1 Sampling Uncertainty

One way to reduce sampling uncertainty is by looking at data availability. The approach is designed to continuously monitor the markets. Different types of data will be acquired and transformed into a time series format. Firstly, the discussion of the globalised markets – economies all over the world takes place. Secondly, the news sources that are used to extract investor sentiment will be introduced.

Although stock markets only work approximately 4 – 7 hours daily, and considering that each of them is located in a different time zone (Asia across UTC+2 to UTC+12, Europe covers UTC-1 to UTC+3, and America is within UTC-10 to UTC-2), one need to construct an approach that monitors all the stock markets 24 hours a day.

The quantitative data, particularly high-frequency data, which represents the complex markets, tends to be lacking in information when stockholders face irrational fear or unfounded hope and change their usual behaviours (Mügge and Stellinga, 2015; Mian and Sankaraguruswamy, 2012)[65][62], or because of changes in fashion and/or technology (Hardie and MacKenzie, 2007)[42]. A similar study of this kind of behaviour has been mentioned in the economics of climate change (Lovell, 2014)[59].

It was intended to explore market data at different frequencies including annually, monthly, daily and higher-frequency. Nowadays, data sources are comprehensive including equity exchanges (NYSE, OMX etc.), financial analysts (Bloomberg, Thomson Reuters, Moody's Analytics etc.), information websites (Google Finance, Yahoo Finance, etc.) and data aggregators (Datastream, Quandl etc.). It is free to access most low-frequency data and some of the high-frequency data (Google Finance and Yahoo finance) provide ticker data for several stock markets for free. However, real-time quotes are not freely and easily available on the web. Some websites offer one real-time quote at a time, but typically only after one has enrolled in a service and/or signed a complicated legal agreement. Other sites approach the problem differently and show

the streaming delayed data. To the best of the author's knowledge, Google Finance offers access to real-time prices from the New York Stock Exchange (NYSE).

Investor behaviour is expected to be that of a rational person – neither risk-averse nor risk seeking. However, it has been argued that during times of market volatility, the investor behaves irrationally and is driven by sentiment, expressed in terms of unjustified fear or blatant optimism. Sentiment proxies are then used in mathematical equations as external factors for testing the auto-regression of market returns – this leads to an estimation of the impact of a sentiment proxy. To identify investor sentiment, the approach connects to the real-time news flow from different news sources. Typically, one can identify two types of media: legacy and social media. Legacy media is the traditional way that people use to gain information including news agencies, newspapers and television programmes; social media is a newer method which allows users to share their opinions, comments, feelings, emotions and moods by posting messages online.

In financial sentiment analysis, the focus recently has been on sentiment that may be implicit in commentary columns (Tetlock, 2007; Garcia, 2013)[92][40], in filings of company reports (Loughran and McDonald)[58], or in the measure of what is known as social media (Antweiller and Frank)[6]. There are commercially available platforms that tend to relate positive attitude by the so-called "likes" on platforms like Facebook, or by a bag-of-words or decontextualised count of negative/positive words on platforms like Twitter. The specialised sections of digital media, where one can see investor-mentor newsletters or commodity "news", comprise news reports and the blogged views of a specialised commodity. Murphy et al. (2015)[66] worked on crude oil prices using shale-oil related news and comments together with commodity news in financial newspapers have shown differences in negative sentiment in these different text types and the effect on oil prices of these different sources. Kelly and Ahmad (2015)[48], explored the effect of crude oil related news on a key market index (S&P 500), and showed that the source of news and commentary is paramount in composing a reliable sentiment proxy. The news sources are summarised in a so-called *"news cline"* and the components are discussed that comprise this *news cline*. Zhao and Ahmad (2015a, 2015b)[103][104] have interpreted the impact of negative sentiment in news and wires on both Danish and Chinese stock markets. (see Table 2.2).

Table 2.2: **News sources and results summary in sentiment analysis:** *, ** and *** denote values of coefficients' statistical significance at 0.1, 0.05 and 0.01 levels respectively. All negative sentiment impact (coefficients) are in **basis points** and all correlations are in percentage (shown in %). Days of the week is shown in brackets.

Author	Data Period	News Source	Severe Market Downturns	Market	Impact/Correlation (Negative)
Data manually scraped					
Antweiler & Frank (2004)	2000	Yahoo! Finance Raging Bull &WSJ	NA	US	- 0.5 % (1d) ***
Tetlock (2007)	1984-1999	WSJ Opinion	1: 1991	US	-4 (1d) *** 4 (4d) ***
Loughran & McDonald (2011)	1994-2008	10-Ks and 10-K405s	1: 2001	US	- 3.9 % (0d)
Garcia (2013)	1905-2005	NYT Opinion	20	US	-4 (1d)
Data via automated programme					
Yu et al. (2013)	07/2011 -09/2011	Google Blogs BoardRead Twitter Google News	NA	Firms	Blog: -0.4 (1d) Forum: -30 (1d) *** Tweet: 0.3 (1d) News: 30 (1d)
Murphy et al. (2015)	2007-2015	News York Times Financial Times	1: 2009	Oil	-7 (3d) ** -4 (4d) *
Zhao & Ahmad (2015a)	2002-2015	Danish Business Digest Esmerk Danmark News	2: 2001, 2008	Danish	-6 (1d) ** 4 (5d) *
Zhao & Ahmad (2015b)	2000-2015	Xinhua News Agency	2: 2001, 2009	Chinese	-10 (1d) *** 7(2d) ***

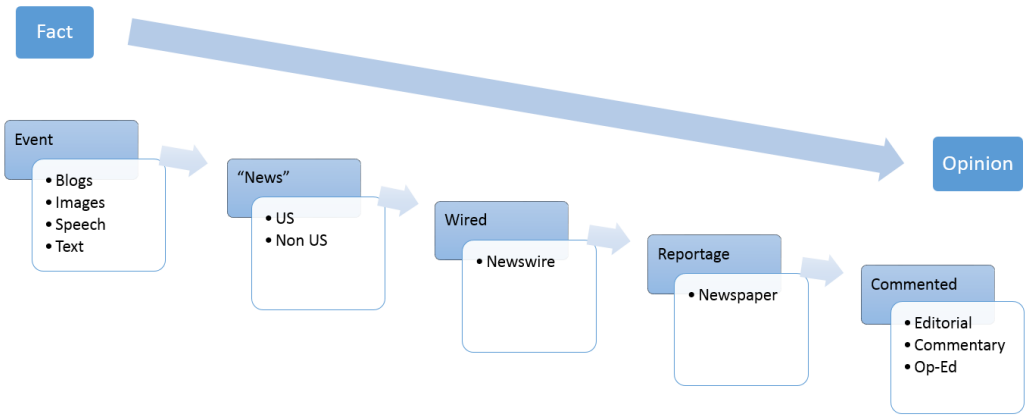


Figure 2.3: **The News Cline**

The *news cline* (Figure 2.3) explains how news travels from when the event occurs to news releases, along with published news reports and peoples' comments. The facts of any event are reported most accurately in blogs, text, speech or images of the event. Seconds or minutes after the event, users of social media post their initial thoughts about the event to platforms like blogs, Facebook or Twitter. Later on, news wire service providers release news reports through wire services (originally telegraphy was used; today the Internet is frequently used). Newspaper organisations source news reports from newswire providers (in fact, nearly every newspaper company is a member of news wire (or wire service)) and then (re-write and) publish the stories after a time-lag of one day. News reports in daily newspapers are usually published on the following day. These news reports still present a high proportion of facts with some background knowledge and explanation added by editors. However, journalists and analysts will make comments and forecasts based on the published facts and typically use some sentiment-laden phrases and opinions into the "editorial" or "commentary" sections in newspapers. All of these commented columns are expressing affect/sentiment in newspapers rather than only stating the fact of the event.

Availability of media data becomes crucial in the case of systems which extract sentiment proxies from these data. There are APIs offered by Facebook and Twitter that allow a system to capture the relevant posts and tweets continuously in real-time. Comparably, there are news aggregators,

such as LexisNexis and ProQuest, which provide daily text contents of news wires and newspapers (including news releases, news articles, op-eds and commentaries). Both the social and legacy media providers mentioned above also offer official translation services.

2.4.2 Modelling Uncertainty

The integration of qualitative and quantitative data has been made plausible by computing systems for acquiring and storing both types of data. A key methodological question has been raised after the data evolution: "Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools." (Varian, 2014)[97].

In this research, the following models are applied to reduce the modelling uncertainty in the approach: regression models to detect quantitative proxies – calendar effects and signalling crash periods; the bag-of-words (BoW) model to extract textual sentiment proxies; Vector Autoregression (VAR) models to integrate quantitative and qualitative data and test the interaction between markets and the sentiment causal impact on stock markets; Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models to measure market volatility and calculate risk evaluations; and machine learning models to help investors in decision-making.

The average value of stock returns has a significant difference according to the day-of-the-week, the month-of-the-year and holidays (Taylor, 2011)[90]. French (1980)[38] analysed the mean and standard deviation for daily Standard & Poor Composite index returns between 1953 and 1977. In general, returns on Mondays have a lower mean value and higher standard deviation than other days. Rozeff and Kinney (1976)[74] states that the US market returns are significantly higher in January just after the New Year than in other months during the period from 1904 to 1974. Furthermore, Ariel (1990)[7] shows that the average returns of days before holidays are 10 times higher than other days during the period from 1963 to 1982. In sentiment analysis, Tetlock (2007)[92] considers the calendar effects by using dummy variables of Mondays and Januarys in his models to control the effects of these abnormal days.

Many researchers, like Schwert (1989)[78] and Taylor (2011)[90], confirm that during recessions, the volatility of returns is unusually high. Evidence shows that, similarly to calendar effects, returns seem to follow patterns rather than go on a random walk during particular periods. During recessionary periods, the news flow about market pessimism increases, and at the same time market trading volume is surprisingly high (Garcia, 2013)[40]. It is assumed that during recessions, markets will be temporarily inefficient, affected by extremely high pessimism sentiment of investors and causing an unusual change in trading behaviours. Previous studies support the notion that markets have irregular movements during recessions; Tetlock (2007)[92] proxies the volatility of returns using VIX index and Garcia (2013)[40] directly measures the impact of news measures during recessions with reference to NBER recession periods.

$$SENTIMENT_T = -0.241CEFD_t + 0.242TURN_{t-1} + 0.253NIOP_t$$

$$+0.257RIPO_{t-1} + 0.112S_t - 0.283P_{t-1}^{D-ND} \tag{2.5}$$

There are so-called indicators showing a market or a company's financial performance. Investors usually look at these indicators to help them making buy/sell calls. Baker and Wurgler (2006)[10] constructed an estimated proxy using a combination of closed-end fund discount ($CEFD_t$), detrended log turnover ($TURN_{t-1}$), the number of IPOs ($NIPO_t$), the first-day return on IPOs ($RIPO_t$), the dividend premium (P_{t-1}^{D-ND}), and the equity share in new issues (S_t) of selected companies (Equation 2.5). They subsequently used this proxy to test the impact of sentiment on 10 equal-weighted stock portfolios comprising different industry sectors and/or firm sizes. Their main finding is that mature/large size stocks are affected by high (positive) sentiment, however, younger/smaller stocks with high volatilities are likely to be affected by low (negative) sentiment. Kumar and Lee (2006)[53] used around two million retail investor transactions over five years in the 1990s to show that small/low price stocks are highly correlated to retail transactions.

Macro-economical factors such as monetary demand and supply, currencies, consumer confidence and GDP/GNP have been analysed in both economics and finance for several decades. Previous studies included macro variables as proxies of sentiment; for example, Siegel (1992)[84] analysed

the 1987 stock market crash in detail with efforts on testing macroeconomic indicators' impact. The chosen indicators, consensus corporate profit forecasts and interest rates were completely unable to explain the stock price movements during the crash. However, Bergman and Roychowdhury (2008)[14] investigated how the Michigan Consumer Confidence Index, a monthly survey which based scores on a linear combination of the responders, affect the companies' disclosure policies. During pessimistic periods, the companies increase the frequency of earnings forecasts, while during optimistic periods, they behave the opposite way. In addition, there are other macroeconomic factors that may affect sentiment analysis in the stock market such as the Purchasing Manager Index (PMI), Dollar Index, Commodity Index and Volatility Index.

Sentiment proxies

The use of quantitative data introduced in the previous section is knowledge driven (e.g. financial reports require accounting skills and technical analysis need quantitative training). In this section, sentiment proxies is discussed which do not require users to have specialised knowledge or training. Text content from news articles and social media messages are easy to understand and the analysis based on these sources are sensible to non-professionals.

There are some noteworthy studies in relation to gathering qualitative sentiment proxies: Cutler et al. (1989)[28] is one of the earliest studies involved in textual sentiment analysis. They found that macroeconomic news is difficult to explain stock market price movements. Andersen et al. (2002)[5] found that the market reacts to macroeconomic realisations (announcements) in an asymmetric fashion. As discussed above, Tetlock (2007)[92] generated his sentiment proxy from the word count of negative and weak affect words in a comment column in the *Wall Street Journal*. Then, using principal components factor analysis, he generated a media pessimism factor. He tested the media pessimism against market price and found that a high pessimism measure leads to price decreasing. He also concluded that unusually high or low pessimism measures result in high trading volume. Garcia (2013)[40] counted the *affect*[5] words of two commentary columns in the *New York Times*, over a period of 100 years (1905-2005) and extracted negative

[5]Emotion or desire as influencing behaviour.

polarity words from the columns and used the frequency as the proxy for investor sentiment. Garcia showed that negative sentiment has a greater impact during recessionary periods (12 basis points) when compared to expansionary periods (3.5 basis points). Ahmad et al. (2015)[2] built a corpus comprising over 5 million news articles. They selected and analysed 20 large US firms over a 10-year period. Distinguishing between English language news and newswire, they employed time-varying regression models and concluded that sentiment analysis on firm-level returns can sometimes be relevant.

Use of multiple models

It is not stable to rely on a single statistical model (noted over 50 years ago in the analysis of airline safety (Barnard, 1963)[12] and in the recent 2008 financial crisis that affected the European markets (Cuaresma et al., 2014)[27]). Having discussed the methods of getting investor sentiment proxies, it has been started looking at how to merge the financial and sentiment data and measure their interaction. Most of the previously mentioned studies applied statistical models to calculate the impact of sentiment and discussed these models in greater depth now.

Statisticians such as Udny Yule (1922)[102] and Maurice Kendall (1976)[49] started studying equities and commodities using regression models. Autoregressive models, where the output variable depends linearly on its own previous values and on an error term, have been used to explore fluctuations in a variety of time series, especially in economics and finance; usually, this dependence is linear in nature. Any deviation from the model can be attributed to a number of causes: for example, there is a steady growth in the value of stock prices or there may be events that are external to the approach that may have a disruptive influence on the values of stock prices – this disruption may be due to changing fashions, market sentiment, extreme weather, disasters or discoveries of many kinds.

Charles Sims (1980)[86] extended the scope of regression analysis considerably by specifying a framework where a regressand (say, today's stock price return) may be regressed against its past values and against any number of other regressors – including traded volume and indeed sentiment. Sims's framework[86] allows us to look at the simultaneous, interdependent analysis of

how a sentiment variable is influenced by the changes in an equity time series. This kind of analysis helps to answer basic questions of causality. There are statistical tests of the directionality of causality, attributable to Clive Granger (see Granger and Newbold, 1977)[41].

Market volatility probably plays another role in the interaction between these markets and there are some studies that have looked at market volatility. Schwert (1989)[78] examined the monthly stock returns data from 1857 to 1987 and found that the stock market is relatively volatile during recessions (e.g. Great Depression) but not everywhere. Koutmos and Booth (1995)[52] studied the transmission mechanism in stock markets. They found that before and after the recession periods, markets are more volatile than average and that the arrival of bad news and the interactions among markets are increased. Kearney and Patton (2000)[46] used conditional variance (GARCH model) as their volatility measure and observed that among French franc, German mark, Italian lira and the European Currency Unit, the German mark plays the main role in volatility transmission.

In addition to statistical/econometric methods researchers have used machine learning techniques like Naive Bayes, Support Vector Machines, Classifier ensembles and neural networks to train a system to learn investor behaviours in conjunction with their text content analysis results (Antweiler and Frank, 2004; Schumaker et al., 2012; Das and Chen, 2007; Bollen et al., 2011; and Yu et al., 2013 [6][77][31][20][100]). A summary of the various econometrics and machine learning methods are summarised in Table 2.3. The approach will choose which model to use.

2.5 Empirical evidence

One of the most popular theories in finance is the Efficient Markets Hypothesis, also known as the Random Walk Theory. The assumption is that if the market is efficient (Fama, 1970)[35], share price movements (returns), which follow a random walk, always incorporate and reflect all relevant information in the price series. In an "efficient market", none of the techniques which are trying to identify the undervalued securities and gain profits from them are effective, and therefore, it is not possible to outperform the market. Most empirical evidence supports

Table 2.3: **Summary of models used in sentiment analysis**

Text type	Source	Method	Estimation	Reference
Online messages	Message Boards	BoW	Naive Bayes Support Vector Machine Classifier ensemble	(Antweiler and Frank, 2004)[6] (Das and Chen, 2007)[31]
Firm releases	EDGAR, Compustat	BoW	Panel Regression Naive Bayes,	(Henry, 2005) (Li, 2010)[56]
Financial News	Wall Street Journal NY Times Dow Jones News News Wires	BoW	OLS & Fama-Macbeth Multivariate regression, VAR & OLS regression Panel & Fama-Macbeth Support Vector Regression OLS Regression	(Loughran&McDonald, 2011)[58] (Jegadeesh and Wu, 2013)[44] (Tetlock, 2007 & 2008)[92][93] (Engelberg&Parsons, 2011)[32] (Schumaker et al., 2012)[77] (Garcia, 2013)[40]
Social Media	Twitter Twitter (+Google)	BoW OpinionFinder	Fuzzy neural network, Naive Bayes	(Bollen et al., 2011)[20] (Yu et al., 2013)[100]
General news	Bloomberg News	Latent Dirichlet Allocation	Linear Regression	(Jin et al., 2013)[45]

the idea that the market is weak or semi-strong form efficient. The approach will first test the market efficiency using the stock returns before taking into account any external factors. This step happens in the modelling process to ensure the correct model, with relevant data, is being used to do the analysis. Some case studies have been conducted during the research to support the approach design. Below a brief introduction is given to some of the tests used in these case studies, specifically to market correlations, calendar effects and sentiment matters.

As mentioned, stock returns are assumed to follow a random walk – that is that stock returns should be independent of each other and have no significant serial correlations and inter-correlations (Taylor, 2011)[90]. The correlations have been tested between market and firm stock returns, the volatility index (VIX) and macroeconomic indicators. Most indices are correlated to each other (with lagged returns) and firms listed on the NYSE are highly correlated to US indices. The volatility index shows a significant negative correlation with most index and firm returns (except Shanghai stock index). The Dollar index unsurprisingly has negative correlations with indices and most exporters (with the exception of positive correlation with Wal-Mart who is a goods importer from all over the world). The purchasing index correlates to indices and most producers (with the exception of one service provider, one pharmaceutical company and one distributor). The commodity index generally correlates to oil companies and energy consumers.

One aim of the approach in the pre-processing stage is to select the correct variables to put into analytic models. If time series are highly correlated (either positively or negatively), it means that some information in these time series is overlapped. This overlapping will discount the forecasting power of the approach. Therefore, the approach decides whether or not to use variables when they are correlated. For example, the Dow Jones Industrial Average (DJIA) index is highly correlated to the Standard & Poor 500 (SP500), Financial Times Stock Exchange 100 (FTSE) and OMX Copenhagen (OMXC) indices perhaps because of close relationships between each country and the US leading effects in the world. However, the DJIA is mildly correlated with the Shanghai Stock Exchange Composite (SHCOMP) index (4%). The user may decide to include the SHCOMP index in the model to test any interaction between markets. However, if the time zone difference is taken into account, the US market does not start their trading

Table 2.4: **Market correlations (with 1^{st} and 2^{nd} lags)**

	DJIA	SP500	FTSE	OMXC	SHCOMP
SP500	97%				
FTSE	53%	54%			
OMXC	39%	40%	68%		
SHCOMP	4%	4%	10%	12%	
DJIA_1	-9%	-9%	26%	31%	11%
SP500_1	-9%	-9%	27%	32%	12%
FTSE_1	-4%	-4%	-5%	8%	12%
OMXC_1	-5%	-4%	-5%	4%	11%
SHCOMP_1	0%	-1%	-4%	-3%	1%
DJIA_2	-4%	-5%	-5%	-2%	2%
SP500_2	-4%	-4%	-5%	-2%	1%
FTSE_2	-4%	-5%	-5%	-2%	-4%
OMXC_2	-4%	-5%	-3%	-3%	-3%
SHCOMP_2	-2%	-2%	0%	1%	-2%

until after the closing bell of the Chinese market. Therefore, it is possible that the information from the US market will be enclosed into tomorrow's Shanghai index prices. The approach will test lagged correlations between markets to ensure that selection of variables is sensible. (Table 2.4 shows market correlations with 1^{st} and 2^{nd} lags[6]; detailed correlations and analysis are in Chapter 4).

2.5.1 Calendar Effects

Day-of-the-week, month-of-the-year and holidays are significant factors in most empirical studies and dummy variables are included in the models to account for these calendar effects. The data sample period in the case studies is 2000 to 2015. During this period, the January dummy variable is significant at the 0.1 significance level. The Monday dummy variable is not significant, perhaps because of the inclusion of the volatility proxies which account for the past 5 days unusual movements of prices. This information could have incorporated the day-of-the-week effects. The

[6] 1^{st} (or 2^{nd}) lag denotes that the time base of the observations in the lagged return time series is shifted back by one day (or two days) compare to the original return time series.

standard deviations[7] of returns of different indices are plotted in Table 2.5 to show that returns on Mondays are more volatile.

Table 2.5: **Calendar effects:** day-of-the-week standard deviations

	DJIA	**SP500**	**FTSE**	**OMXC**	**SHCOMP**
Monday	1.30	1.39	1.32	1.44	1.93
Tuesday	1.19	1.31	1.17	1.24	1.50
Wednesday	1.16	1.24	1.20	1.34	1.53
Thursday	1.21	1.28	1.20	1.25	1.55
Friday	1.07	1.11	1.15	1.23	1.44

2.5.2 Sentiment Matters

Sentiment analysis has been popular in the past two decades, and there has been much research on extracting sentiment from texts. Antweiler and Frank (2004)[6], Tetlock (2007)[92], Loughran and McDonald (2011)[58], Garcia (2013)[40], Yu et al. (2013)[100], Murphy et al. (2015)[66], Zhao and Ahmad (2015a, 2015b)'s[103][104] recent findings use news sources from blogs, Twitter tweets, Facebook posts, news wires, news reports, ed-ops, and commentaries. The impact of sentiment proxied from these news sources has proved significant on financial markets. Some case studies have been carried out involving US, Danish and Chinese markets (Table 2.6), and found that negative sentiment has a significant impact on indices that represent these markets. Generally, if the negative sentiment increases by one standard deviation, the index returns will drop 5 to 10 basis points over one day. However, after the initial negative impact, the markets generally recover (albeit in different time periods). The Chinese market recovers after one day, and the US and Danish markets recover after four to five days[8].

[7]The difference between the returns on each day of the week and the average of the returns in the sample period. The high the standard deviation the more volatile return a stock has on a day.

[8]A detailed analysis is in Chapter 4

Table 2.6: **Sentiment impact: negative tone on different markets**

		Dependent variable: $return_t$					
	Coef.	**DJIA**		**OMXC**		**SHCOMP**	
Negative	$t-1$	-4.8	***	-6.0	**	-9.5	***
	$t-2$	2.2		0.9		7	***
	$t-3$	-1.0		0.6		-0.5	
	$t-4$	4.9	***	-0.7		1.8	
	$t-5$	2.5		4.4	*	-0.6	
Negative	$\chi^2(5)$**[joint]**	20.20	***	5.13		15.50	***

2.6 Visualising Investor Sentiment

Previous studies have commonly used regression models (including simple regression, Vector Autoregression, GARCH, etc.) to compute investor sentiment analysis. In regression models, a coefficient of any variable denotes the correlation of movements between dependent and independent variables in standard deviations. Therefore, one of the most popular ways to interpret the impact of investor sentiment is to show how much change in basis points an independent variable causes (Tetlock, 2007; Garcia, 2013)[92][40]. Conditional volatility is computed using the GARCH model to estimate the stochastic risk scores of the stock returns (Taylor, 1994)[91]. Recently, rolling regression (regressions with rolling windows, say 1-year moving regression) has been used in representing the periodical impact of any independent variable in a rolling window (Murphy et al., 2015; Ahmad et al., 2015)[66][2]. With this technique, the model shows the year on year impact (using previous year's empirical evidence as the current year's reference) of investor sentiment on stock markets.

It has been argued that "the visual information on a plot can be greatly enhanced [...] by computing and plotting smoothed points" (Cleveland, 1981)[24] under the rubric of robust locally weighted regression that essentially does piece-wise regression.

A time-ordered plot of the residuals obtained from the Equation 2.4b in Table 2.1 and the newspaper sourced sentiment shows two noisy time series superposed on another (Figure 2.4a). A relationship between the two time series was expected to be depicted, however, is not clearly

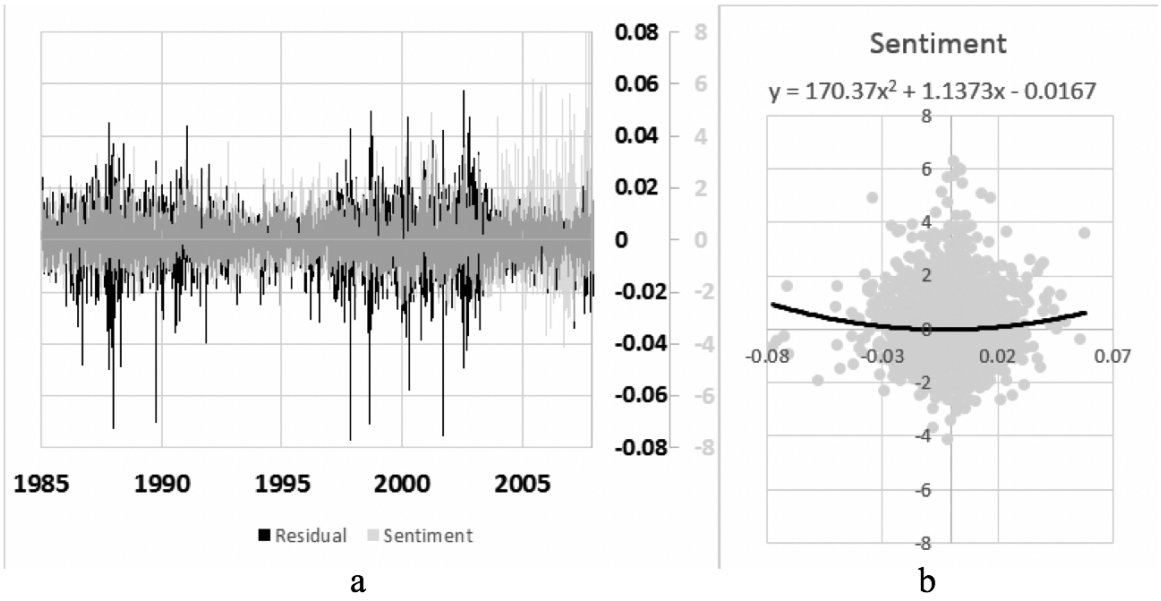


Figure 2.4: **a (left): time varying behaviour of residuals as extracted from Eq 1b and sentiment harvested from WSJ AofM; b (right): relationship between sentiment and residuals using DJIA and WSJ between 1984 and 2007**

shown on a simple plotting figure. If sentiment is plotted against the residual and then do a simple regression on the two, one sees a sort of relationship between them (Figure 2.4b) possibly obscured.

2.7 Summary

Investor behaviour is expected to be that of a rational person neither unduly risk averse nor excessively risk seeking. However, it has been argued that during times of market volatility, the investor behaves irrationally and is driven by sentiment, expressed in terms of unjustified fear or blatant optimism.

Previous studies have been summarised in data mining, financial econometrics and sentiment analysis. To reduce sampling uncertainty, quantitative and qualitative data should be obtained from the globalised markets in real-time. Researchers measure quantitative data using financial reports, market prices and economic indicators and qualitative data using sentiment proxies,

especially the sentiment extracted from textual sources. Some useful models have also been described in data analysis - simple regression, vector autoregression, GARCH and machine learning. With multi-model analysis, researchers are able to test the impact of multi-variables (quantitative and qualitative sentiments) on the target variable (stock market data). Market volatility has been discussed where researchers have been using it as a transmission factor to test the interaction between different markets. LOWESS model has been used to analyse the relationship between return residuals and negative sentiment and these models serve as a visualisation tool of investor sentiment. In Chapter 3, the methods in use will be discussed in order to build the approach.

Chapter 3

Methods

3.1 Introduction

This chapter outlines the method developed in this thesis. The method achieves the following three tasks:

1. Data harvesting
2. Sentiment analysing
3. Relationship visualising

The approach will collect a variety of financial and textual data from data providers and aggregators. A Multi-model design will allow the approach to choose the best technique to support the user's decision making.

The approach requires quantitative and qualitative data as input in the form of both historical and real-time data. One use historical data to calculate the results that help detect risk-level in the real-time analysing process. In practice, the approach analyses live data and generates a sort of risk indicator to alert investors that the market is facing unexpected movements.

3.1.1 Quantitative Data

All macroeconomic indicators, market performance ratios, and stock price movements are considered to be quantitative data. A number of data providers (or aggregators) provide access to financial data, which is composed of different types (financial statements, economic indicator announcements, stock prices, etc.) and frequencies (annually, quarterly, monthly, daily and high-frequency). The approach collects economic indicators and stock prices as time-series in different frequencies.

3.1.2 Qualitative Data

The investor sentiment is articulated through news and opinions that are made available to other investors, either as the report of events related to an unexpected market downturn or upturn or the expectation of such events. The former is typically published in newspapers or online media as reportage, and expectation is usually in opinion columns. The approach collects historical news articles from some major content providers (i.e., LexisNexis and ProQuest) and real-time news from live information providers (i.e., Twitter tweets, Facebook posts, RSS feeds and newspaper websites).

3.2 Autocorrelation of Returns

3.2.1 Linear Model

A linear regression model is used to explore the relationship between a dependent variable and an independent variable. Scientifically, a regression model helps one to understand how a dependent variable changes when the independent variable is varied.

In this case, a linear regression model relates a time series Y_t to a function of X_t and β:

$$Y_t = \alpha + \beta X_t + \epsilon, t = 1, \ldots, T \tag{3.1}$$

where the unknown parameter, denoted as β, may represent a scalar or a vector, the independent variable is denoted by X and the dependent variable is denoted by Y.

Vector Autoregression Model

In econometrics, and to a lesser extent in finance, the tendency until the late 1980s, was to use regression for a key variable, for example, gross domestic product or share price return, to estimate the impact of the variable. Charles Sims pointed out this focus on a simple variable does not reflect an economy or a market - where there are many other independent variables (Sims, 1980)[86]. Charles Sims proposed the vector autoregression (VAR) - the 'vector' was a vector of key variables: GDP plus import/export data, a share price, market holiday variable, and so on.

Similar to simple regression, Vector Autoregression (multiple regression) has the following input:

$$\begin{gathered}((x1)_1, (x2)_1, (x3)_1, ..., (xK)_1, Y_1 \\ ((x1)_2, (x2)_2, (x3)_2, ..., (xK)_2, Y_2 \\ ((x1)_3, (x2)_3, (x3)_3, ..., (xK)_3, Y_3 \\ ..., \\ ((x1)_n, (x2)_n, (x3)_n, ..., (xK)_n, Y_n\end{gathered} \tag{3.2}$$

where variable Y is defined as the "dependent variable." There are many independent variables in VAR denoted by x.

The output from a VAR is also a "fitted regression model". A VAR model implies that Y is a linear function of the predictors (with n different lags of each predictor), plus statistical noise:

$$Y_i = \beta_0 + \beta_1 L_n x1_i + \beta_2 L_n x2_i + \beta_3 L_n x3_i + ... + \beta_k L_n xK_i + \epsilon_i \tag{3.3}$$

Many statistical summaries can also be produced including R^2, standard error of estimate, t statistics for the β's, an F statistic for the whole regression and so on ...

The coefficients (the β's) are not random but unknown quantities. The noise terms ϵ_i are random and unobserved. The assumption these errors are that they are statistically independent with a mean (μ) equal to 0 and a standard deviation (σ) equal to 1. One look at the significance of coefficients and $F-$ statistics[1] to discover the predictive information of the independent variables.

3.2.2 Volatility Model

Conditional volatility, together with the stylised facts of returns, helps indicate whether a market is efficient. In simple regression models, it is assumed that the variance of the error term remains constant over time. It has been suggested that most of the key indices, firm stocks and commodities show a degree of volatility clustering: high variance during extreme periods and low variance during normal periods. In order to understand the extent to which the returns are impacted by volatility, generalised autoregressive conditional heteroskedasticity (GARCH) models are typically used to compute the conditional variance (h_t). The GARCH models is used for generating the variance of the error terms - h_t (Bollerslev, 1986; Taylor, 1986)[21][89]:

$$h_t = \omega + \alpha^{'}(r_{t-1} - \mu)^2 + \beta^{'} h_{t-1} \tag{3.4}$$

The model produces h_t, the conditional variance, a one-period ahead estimate based on the past standardised return r_t. This equation explains high volatility begetting high volatility and vice versa: $\alpha^{'}$ contains the information about asset risk during the previous period and $\beta^{'}$ interprets dependency on variance during the previous period for the daily log returns.

[1]The ratio of the mean regression sum of squares divided by the mean error sum of squares. The value of Prob(F) is the probability that the null hypothesis for the full model is true.

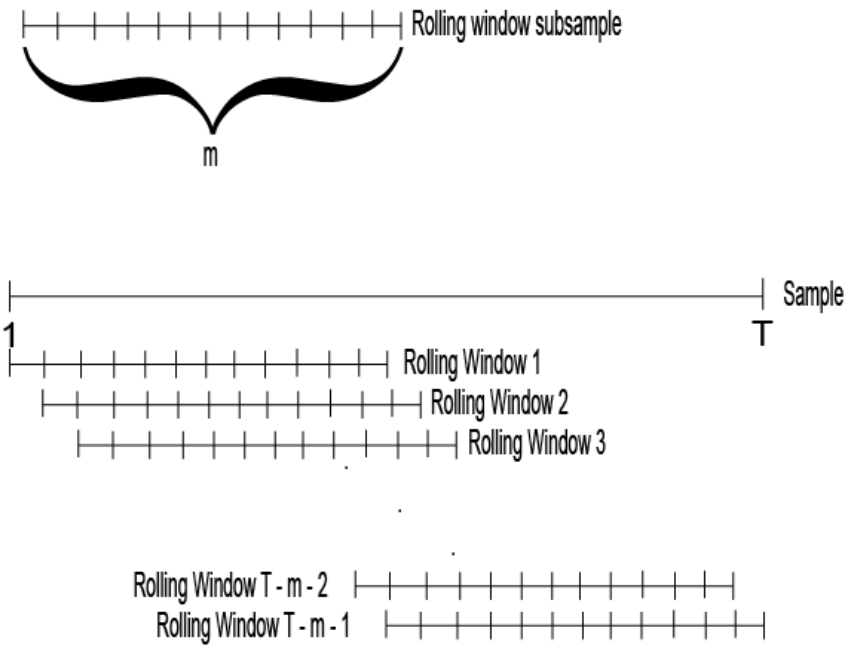


Figure 3.1: **Rolling regression illustration**

3.2.3 Rolling Model

Suppose that one has data for all periods in the sample. To back-test the data is to check the predictive performance of several time intervals using a rolling window.

First, it is important to choose a rolling window size, m, i.e., the number of consecutive observations per rolling window. The size of the rolling window depends on the sample size, T, and periodicity of the data. In general, a one-year (250 days) rolling window is used for daily frequency financial data. If the number of increments between successive rolling windows is 1 period, then the entire data set is partitioned into N = T m + 1. The first rolling window contains observations for period 1 through m, the second rolling window contains observations for the period 2 through m + 1, and so on. In the one-year rolling window case, the first rolling window contains observations for period 1 through 250 and the second contains observations for the period 2 through 251. The Figure 3.1 illustrates the partitions.

3.3 Content Analysis

A text corpus is downloaded from multiple sources including news agencies, newspapers, blogs, editorials and commentaries. The affect analysis approach stores the collection of the captured news which, in turn, can be used to conduct a historical analysis over a selected period of time. Affect words are used to express the evaluation of an event with negative or positive sentiment; affect words are used to estimate a person on being active or passive in a situation; and affect words are used to express depth of emotion - strong or weak emotion. Affect words are used in human relationships and many other instances where one express emotion, activity, human ethical value. Stone et al. (1966)[87] extended the use of such tagged glossary to texts in sociology, human relations and many other areas. They created a system called the General Inquirer (GI)[2] and the dictionary used in GI is referred to as GI dictionary. Many studies have used machine learning methods in sentiment analysis using textual data. In supervised learning, the main part of the learning algorithm is the use of tagged examples, often referred to as "training sets," which are distinguished from a separate set of tagged examples "test sets" for evaluating the classifier's accuracy. The simpler dictionary-based method, together with the Harvard GI dictionary, is used to extract sentiment term frequencies from the corpus. The sentiment time series are used to construct the investor sentiment proxies. The following section introduces the models to be applied to proxy the investor sentiment.

3.3.1 Dictionary-based Analysis - General Inquirer

The most popular word list was originally developed by a group of social psychologists at Harvard University in the 1960s. The primary goal in this section of the thesis is to provide a computational method for content analysis, defined by the author as any research technique that is reasoned in a systematic and objective way to identify particular features in the text. The General Inquirer (GI) used a built-in dictionary (Stone et al., 1966)[87]. In this dictionary, words are mapped to one or more content categories so that semantically identical words will

[2] A computer-assisted approach for content analyses of textual data.

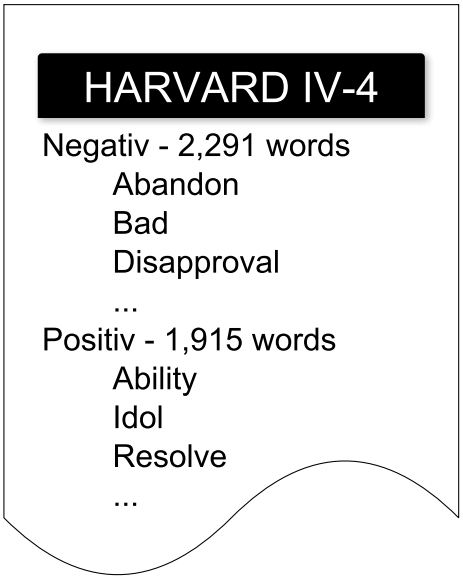


Figure 3.2: **An example of two categories** - Negativ and Positiv in Harvard IV-4 dictionary

be represented by the same concept. For each word in the input text, the system performs a dictionary lookup of the word and adds any applicable category for each word. The Harvard IV-4 Dictionary (built-in for GI program) contains semantic categories from a variety of fields, including economic, political, legal, military, and emotion (e.g. happiness and painfulness). For this study, it is essential that the impact on assessment, activity and effectiveness be represented by positive and negative, active and passive as well as strong and weak GI categories, respectively. After many years of construction, GI dictionaries have been revised several times to add or delete new categories, and subsequently to revise terms belonging to existing categories. The dictionary contains over 11,000 words tagged with an ontology comprising 182 categories: valence, semantics, emotions, language institutions, social words, places and objects, and etc. The two large sentiment outlook categories, positive and negative, comprise of 1,915 and 2291 words respectively (see Figure 3.2).

Other scholars have also studied and extracted dictionaries that are specific to the field of financial disclosure (e.g. Loughran and McDonald (2011)[58]). These dictionaries analyse textual sentiment more accurately than GI in specific contexts because GI is created for general environ-

mental languages rather than special environmental languages such as the financial environment. However, in the analysis of text sentiment, this thesis focuses on understanding the general psychological changes of all investors in the market, rather than the professional understanding of individual practitioners. Therefore, this study argues that GI, as the foundation of this word list in psychology, is more suitable for the general investor's sentiment analysis.

3.3.2 Bag-of-Words (BoW) Model

Economists and finance experts have been using content analysis to extract media sentiment, specifically a BoW model for extracting investor sentiment from a text corpus. Essentially a BoW model assumes that words are distributed independently of each other in a text. An affect analysis system based on the BoW model is used to identify individual word frequency. These words are matched against a pre-existing list of words, e.g. in a sentiment "dictionary". It has contributed significantly to the development of investor sentiment analysis in recent years. For analysing investor sentiment, the words in the glossary are further classified into sentiment categories – e.g. categories of positive or negative. Every time the programme finds a word belonging to a given category, the category count is increased correspondingly[3].

The workflow of the analysis is as follows: the quantitative and qualitative data comes in as input to the approach. Researchers formalise the two types of data into times series. The stylised facts (a term used in economics to refer to empirical findings that are so consistent that they are accepted as truth) tell whether the markets are efficient. Then all of this data goes into the regression models and the results show whether each of the variables has a causal effect on the others. Finally, depending on the model results, researchers figure out whether the results reject the hypothesis tests and the conclusions (Figure 3.3).

[3]It is ideal to identify when the sentiment is negative (perhaps ironic) if the word comes with positivity and to solve word-sense disambiguation that may occur technical issues, nevertheless, the method conducted in the thesis disregards the order of words and captures the negativity of the content using relative frequency of the negative words.

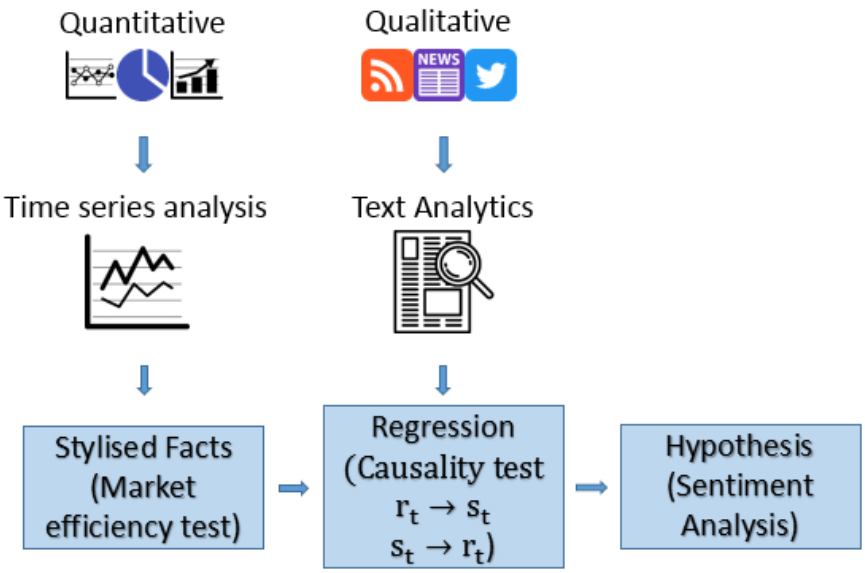


Figure 3.3: **Sentiment analysis work-flow**

3.4 Aggregation of Two Types of Data

3.4.1 Multivariate Vector Autoregression Model

The regression schemes have been displayed in Table 2.1 earlier in Chapter 2. The output from the VAR modes is also a fitted regression model. A VAR model implies that dependent variable Y is a linear function of the predictors (with n different lags of each predictor), plus statistical noise. It has been designed to test the influences between internal (5 lags of the dependent variable) and external variables (5 lags of quantitative and qualitative variables) within five working days.

Many models have been discussed and two different major types of data. The approach will integrate different data and models. A four-step data selection strategy (Table 3.1) has been designed to test the internal (5 lags of the dependent variable) and external variables (5 lags of quantitative and qualitative variables). The first step mainly tests the autocorrelation of the dependent variable with its lagged values (returns). The second step adds the lagged values of an external constraint as an independent variable. It shows the impact of key external quantitative

constraints on the regressand. The third step tests the lagged values of the qualitative variable and checks the impact of it. The fourth step finally put the quantitative and qualitative variables in the equation together and checks the interactions between them and the regressand.

Following the practice in econometrics and in finance, a linear relationship is posited between these different variables and their historical values. This model is a simplification of what happens in dynamic systems like financial markets, so there will be a residual error[4]. The expectation here is that the error is independently and identically distributed ($i.i.d$) (Heteroskedasticity and autocorrelation in the residuals are dealt with using HC robust standard errors. (Newey and West, 1987)[67]). The endogenous variables to be used include five lags of indices returns and negative sentiment, to control for autocorrelation. Exogenous variables include the index volatility proxies and dummy variables for bank holidays, extreme events[5], day-of-the-week and month-of-the-year effects. The 5^{th} order Vector Autoregressive model with the error term ϵ_t is defined. Four tests are designed to evaluate the quantitative and qualitative constraint impacts step by step (Table 3.1).

In the four-step analysis, the approach selects necessary variables from correlation tables; this process ensures that each variable is independent of others and prevents serial correlation and inter-correlation.

Figure 3.4 shows an architecture diagram of the functionality of the text preprocessing and content analysis that is designed for the sentiment analysis approach.

3.4.2 Semi-parametric Estimation of Sentiment on Return - Robust Locally Weighted Regression

A time-ordered plot of the residuals is obtained from the Step IV in Table 3.1. If sentiment against the residual via the simple linear regression is plotted, the results do not clearly show the relationship between the two (Zhao, Kelly and Ahmad, 2017)[105].

[4]The residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean).

[5]i.e. The Black Monday on 19 October 1987.

Table 3.1: **Regressand(Response/explained)(Market return):** $L_n\beta(x_t) = \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_n x_{t-n}$

Step	Regressor (predictor/controlled)	Equation (3.5)
I. Internally consistent model	Lagged values of Regressand ($L_5\beta r_t$)	$r_t = \alpha + L_5\beta r_t + Exog_t + \epsilon_t$
II. Check impact of key external quantitative constraint on the regressand	$L_5\beta r_t$ + lagged values of external constraint ($L_5\delta r_t^{ex}$)	$r_t = \alpha + L_5\beta r_t + L_5\delta r_t^{ex} + Exog_t + \epsilon_t$
III. Check impact of key qualitative constraint on the regressand	$L_5\beta r_t$ + lagged values qualitative variable ($L_5\gamma Neg_t$)	$r_t = \alpha + L_5\beta r_t + L_5\gamma Neg_t + Exog_t + \epsilon_t$
IV. Aggregate quantitative and qualitative variables	$L_5\beta r_t + L_5\delta r_t^{ex} + L_5\gamma Neg_t$	$r_t = \alpha + L_5\beta r_t + L_5\delta r_t^{ex} + L_5\gamma Neg_t + Exog_t + \epsilon_t$

One will use a non-linear function - Locally Weighted Scatterplot Smoothing (also known as LOWESS) (Cleveland, 1979)[23] to explore the relationship between regression residuals and sentiment values. Similar to a moving average method, the smoothed value in LOWESS is decided by neighbouring data points defined within a span[6] and is weighted as the regression weighting function and is also determined for the data within the span.

Three components of a local regression smoothing algorithm are:

1. Calculate the weights for each data point in a span:

$$w_i = (1 - |(x - x_i)/d(x)|^3)^3 \tag{3.6}$$

[6]A percentage of the total number of data points in the data set.

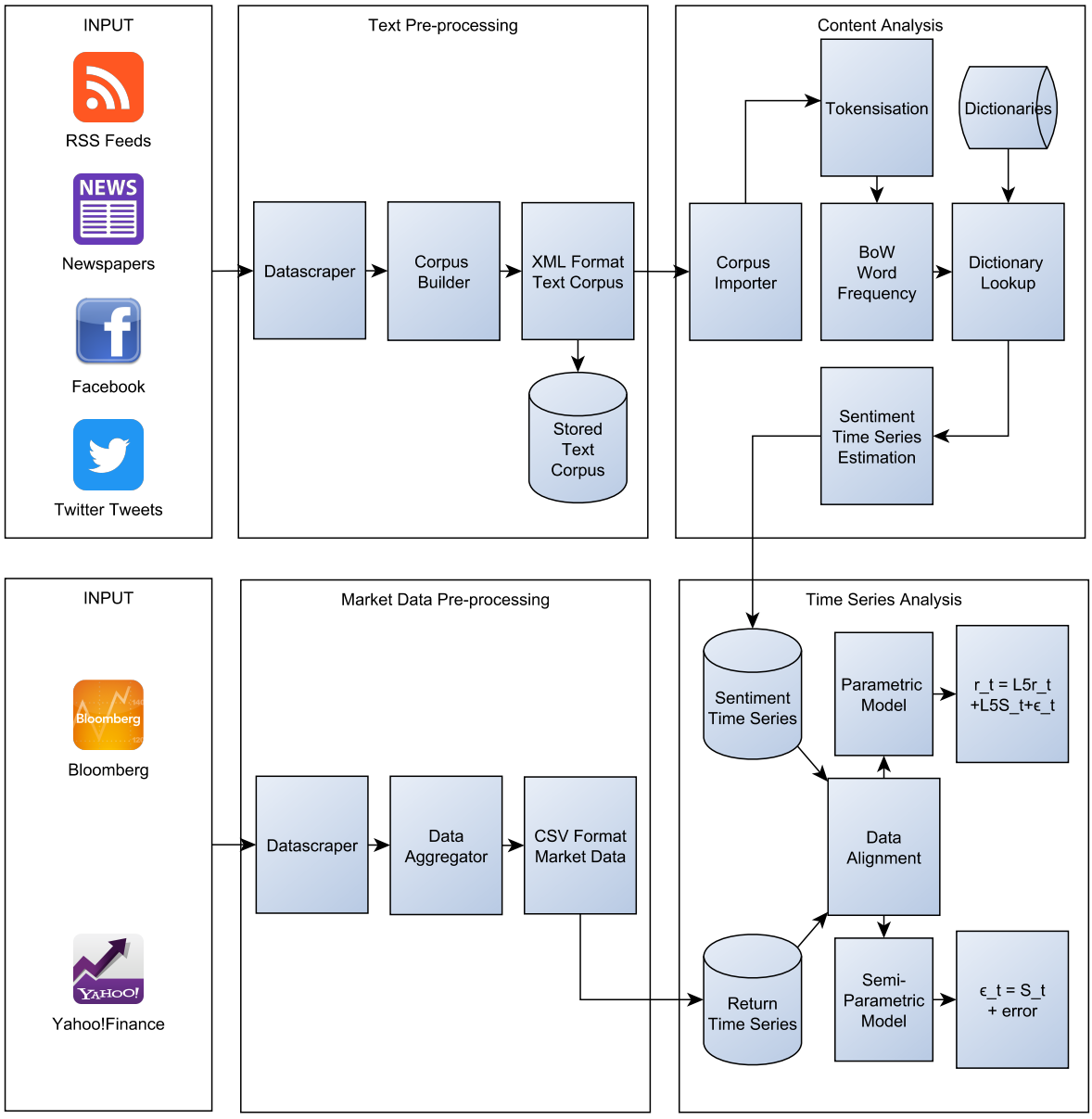


Figure 3.4: **The computational approach diagram**

x is the predictor value associated with the response value to be smoothed, x_i are the nearest neighbours of x as defined by the span, and $d(x)$ is the distance along the abscissa from x to the most distant predictor value within the span.

2. Use OLS for regressing smoothed residuals against smoothed sentiment.

3. Output graph of the smoothed values.

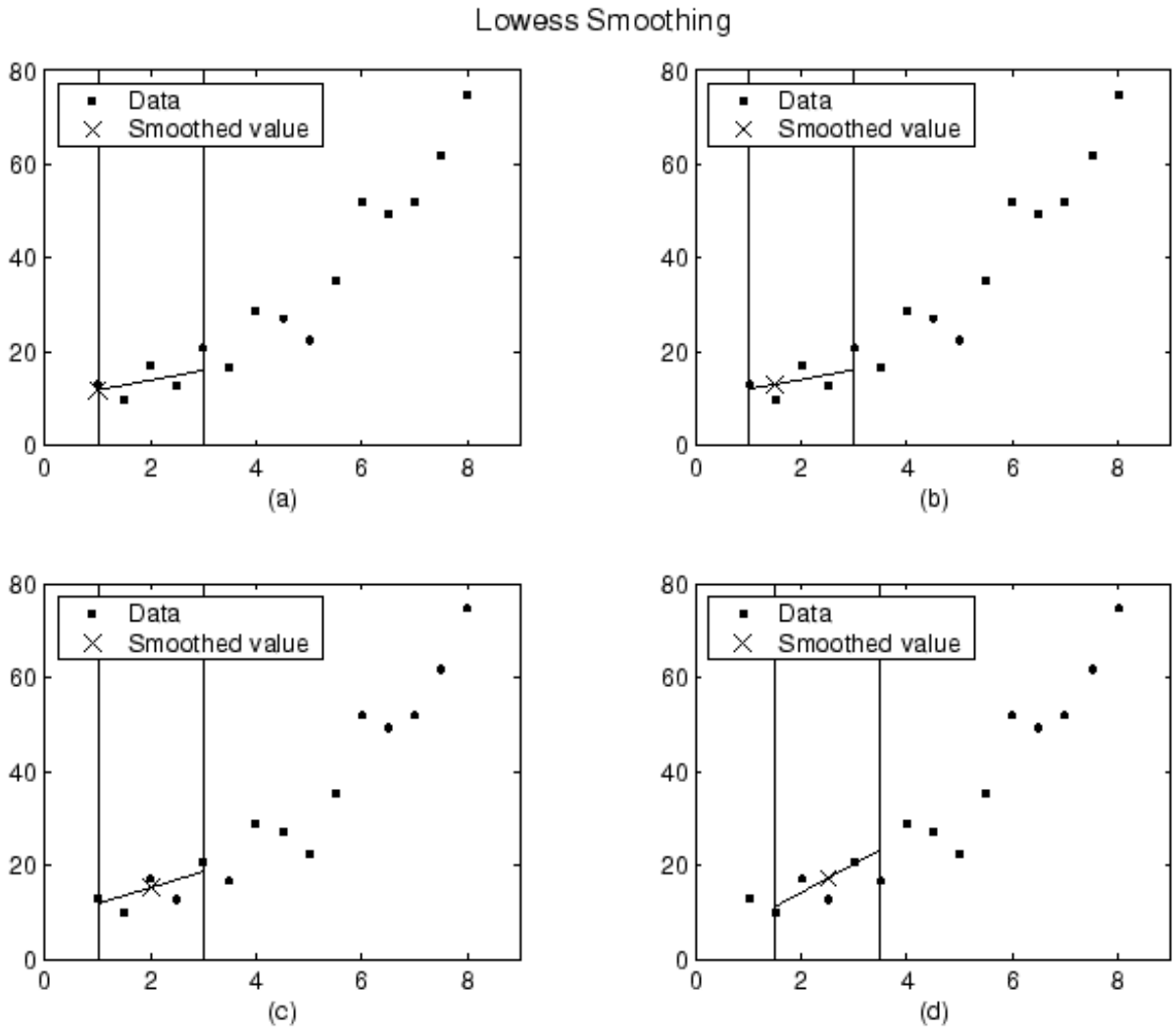


Figure 3.5: **LOWESS smoothing procedures**

The LOWESS smoothing procedures are present in Figure 3.5.

The main reason for using LOWESS is that the local polynomials for each subset of the fit data are almost always first-order or second-order, that is, local linear. Therefore, a zero degree polynomial is used to convert LOWESS to a weighted moving average. Higher-order polynomials are theoretically feasible, but the model does not truly conform to the purpose of LOWESS. LOWESS is used when any function can be well approximated in small neighbourhoods by low-order polynomials, and a simple model can easily fit the data. High polynomials tend to over-fit the data in each subset and are numerically unstable, making accurate calculations difficult.

3.5 Critique of Method and Data

3.5.1 Choice of Texts in the *News Cline*

Many studies have investigated investor sentiment extracted from one or two sources (company filings, news reports, newswire, blogs, editorials and commentaries) (Baker and Wurgler, 2006; Tetlock, 2007; Loughran and McDonald, 2013; Garcia, 2013; Ahmad et al., 2015; etc.)[10][92][58][40][2]. Some work has been done on index-level sentiment analysis and additional work has been conducted at the firm-level. However, a comprehensive comparison will be conducted between all different sources, focusing in particular on firm-level sentiment analysis. The aim is to investigate the impact of sentiment extracted from each particular source and contrast it with the results that are calculated from all sources.

3.5.2 Language Choices

The question is raised at the outset of the thesis that investor sentiment may behave differently in different markets, especially the textual investor sentiment in different languages. So far, all of the research has been conducted using English language corpora to avoid the inconsistency of meaning in different language. In addition, it is argued that language may change over time. As all the text contents were collected from formal media including newspapers and newswires, the language is believed consistent and not to affect the accuracy of the sentiment results in this thesis. However, as discussed in the future work in Chapter 5, if one decides to extend the analysis using social media or in other languages such as Chinese. It is crucial to consider the sensitivity of language in such conditions.

3.5.3 The Use of BoW

The BoW model focuses completely on the words, or sometimes a string of words, and usually pays no attention to the "context" as the order of words in BoW method is less important than the frequency of the word occurrence. The approach only uses two steps with the BoW model.

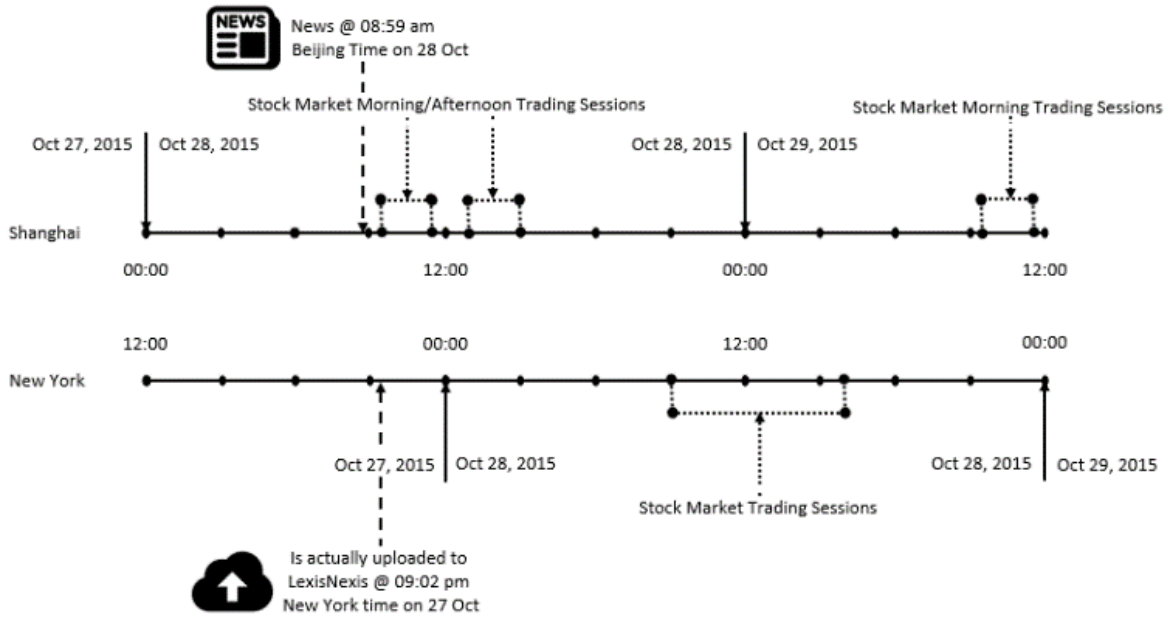


Figure 3.6: **Time stamps in news from *Xinhua news agency***

First, a bag with polarity (positive and negative) words is chosen. The only question here is whether the text content is "black or white". Second, the degree of polarity is measured using the relevant frequency of the positive and negative words that occur in the document. The BoW model has been used by researchers over the past two decades as the de facto standard of financial articles. The advantage of using BoW is that the model provides a suitable proxy of the text content and it is simple to apply.

3.5.4 Consideration of "Time"

Newspaper organisations source news reports from newswire providers and (re-write and) publish the stories. As a consequence, there is a significant gap (taken into account in the calculations) between the time the news is available from wire services (seconds or minutes after the event) and some audiences actually reading it (in the following day's newspaper).

Most news articles that have been uploaded to LexisNexis are one day delayed (lagged). This news was released on the day of the events however was only documented by, say, *Danish Business Digest* a day after. It is also noticed that the date stamp on the news from *Xinhua news agency*

uploaded by LexisNexis was in New York local time, however, the original news was in fact published by Xinhua in Beijing local time according to Xinhua website (Figure 3.6). Similarly, the time stamp on the news from American newspapers (New York Times) and European newspapers (Financial Times) are in different time zones. The mixed time stamps cause inaccurate estimation of the investor sentiment in the manner of timing. A conversion programme for fixing timing errors is used in the research.

3.6 Summary

In this chapter, the methods and techniques have been described that are supportive in building an automatic analysing approach. The approach connects to APIs and automatically collects data. Different data types will be analysed and transformed into the time series format. Both parametric and semi-parametric models will integrate quantitative and qualitative data and compute the sentiment impact analysis.

An integrating approach has been introduced that will systematically collect and analyse data. It will provide visualised results helping to understand the impact of sentiment. In the next Chapter, sample datasets will be collected to test the sentiment analytic approach at both aggregate market- and firm-level through five case studies.

Chapter 4

Evaluation and Case Studies

4.1 Introduction

It is designed to build a systematical approach, and the data and techniques used to identify the impact of sentiment on financial markets will be part of this approach. First of all, whether the sentiment impacts financial markets using the proposed methods will be tested. The approach's aforementioned sentiment and price/return analytic modules will be evaluated through a number of case studies. The first step is to reproduce and extend Tetlock (2007)[92]'s results and then implement the methods into further market studies. Financial and sentiment data will be used for the period Jan 1984 – Jun 2015.

4.2 Data Collection and Pre-processing

To begin with, investor sentiment is analysed in a *Wall Street Journal* opinion column, namely "Abreast of the Market" (AofM). Tetlock (2007)[92] used news articles from AofM covering the period 1984 to 1999, and tested the extracted sentiment on the DJIA index. The same news articles have been collected to confirm that the method will conclude with the similar sentiment influence, and the data period has been extended to investigate the impact of sentiment in different periods from various sources.

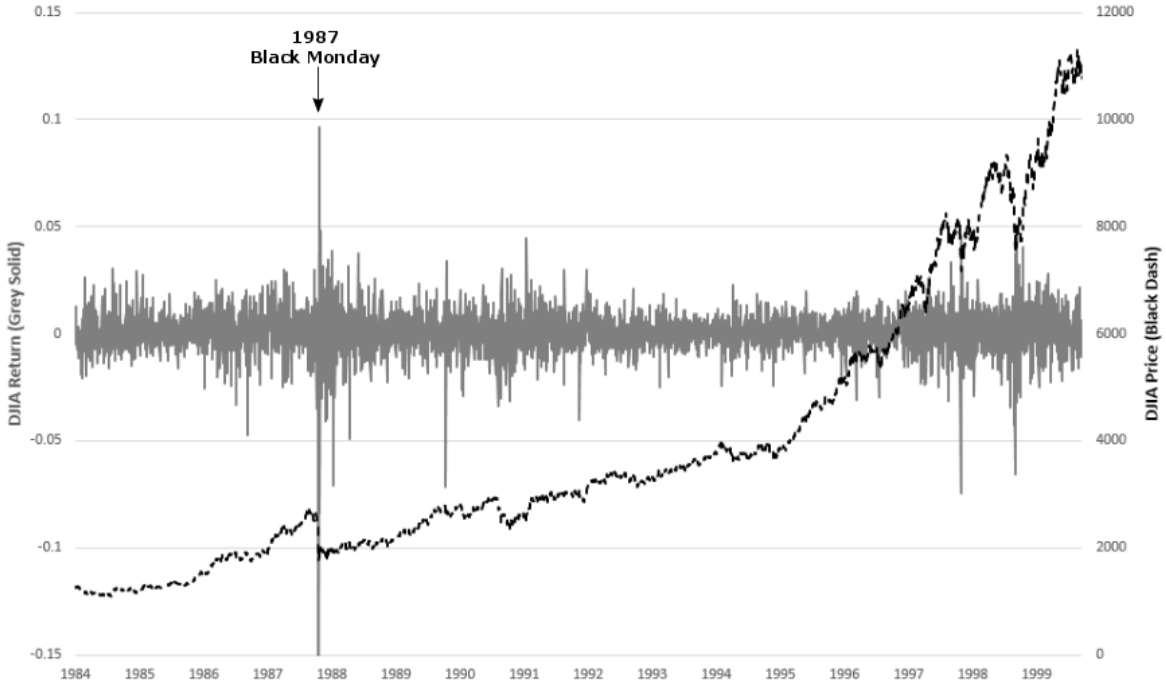


Figure 4.1: **DJIA price and return, 1984-1999**

The further tests have been conducted are on a market that represents EU countries and a market that represents Asian (emerging) countries and on the interaction between these market indices over the last 15 years. News corpora about these two markets are collected following similar criteria so that the results will be comparable.

For comparison, company data has been collected from 23 top Fortune 500 companies, along with relevant U.S. newspaper news texts. The collection of these data is automated however very time-consuming.

4.2.1 Quantitative Data

US - Market-Level Data

The first market to look at is the world's largest national economy, the United States, representing about 20% of global GDP. The U.S. dollar is the most used currency and is the world's foremost reserve currency. NASDAQ Composite, Dow Jones Industrial Average (DJIA) and S&P 500 are

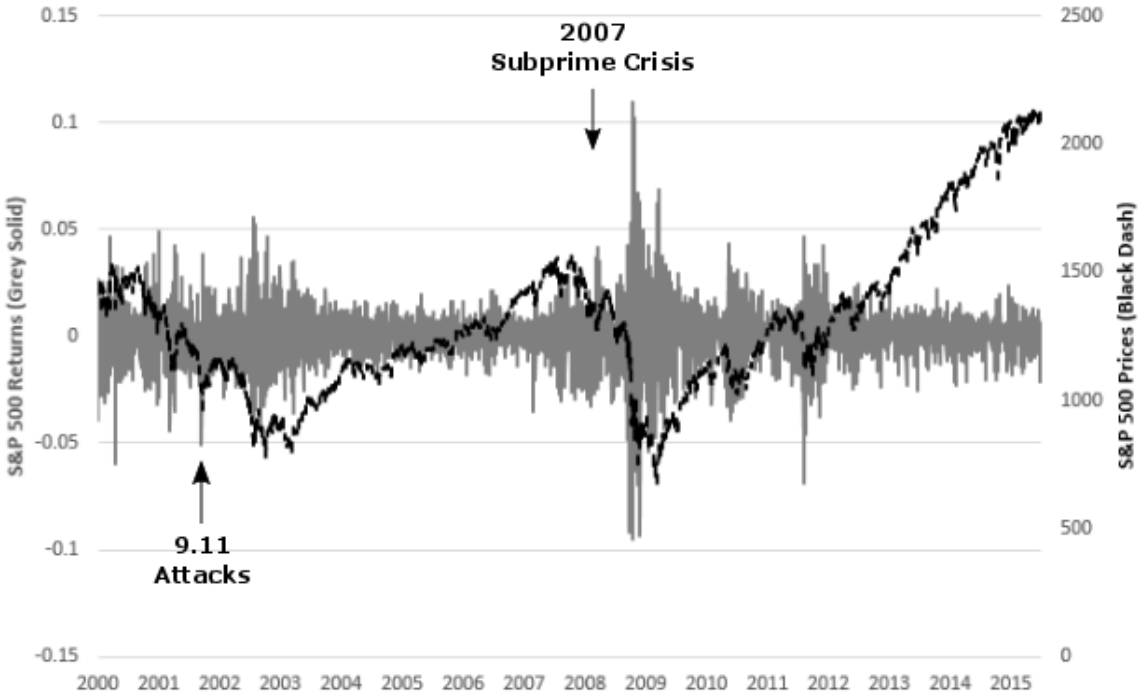


Figure 4.2: **S&P 500 price and return, 2000-2015**

three major U.S. stock indices. Dow Jones Industrial Average (DJIA) and Standard & Poor's 500 (S&P 500) index prices are available since 1898 and 1950 respectively. The DJIA daily prices is used from 1984 to 1999 (Figure 4.1 shows the daily prices and returns of DJIA index) in the computations to reproduce Tetlock's results as the benchmark. In the regression model, the daily volumes of the New York Stock Exchange (NYSE) stock exchange, recorded starting from 1980, is used as the measurement of trading behaviour. In addition, the S&P 500 index (Figure 4.2) is used as the external constraint in the Danish and Chinese market case studies.

US - Firm-Level Data

From Fortune 500 Global, 23 large firms[1] has been chosen. Stock market data of these firms are available starting from 01/01/2000 until 30/06/2015. 19 out of the 23 companies are listed on the NYSE and four of them are listed on NASDAQ. Five of them are non-US based companies

[1]Apple, AT&T, Boeing, BP, Chevron, CISCO, Conoco Phillips, Exxon Mobil, Ford Motor, General Electric, Home Depot, HP, IBM, Intel, Johnson&Johnson, Merck, Microsoft, Pfizer, Shell, Total, Toyota, Verizon, and Wal-Mart.

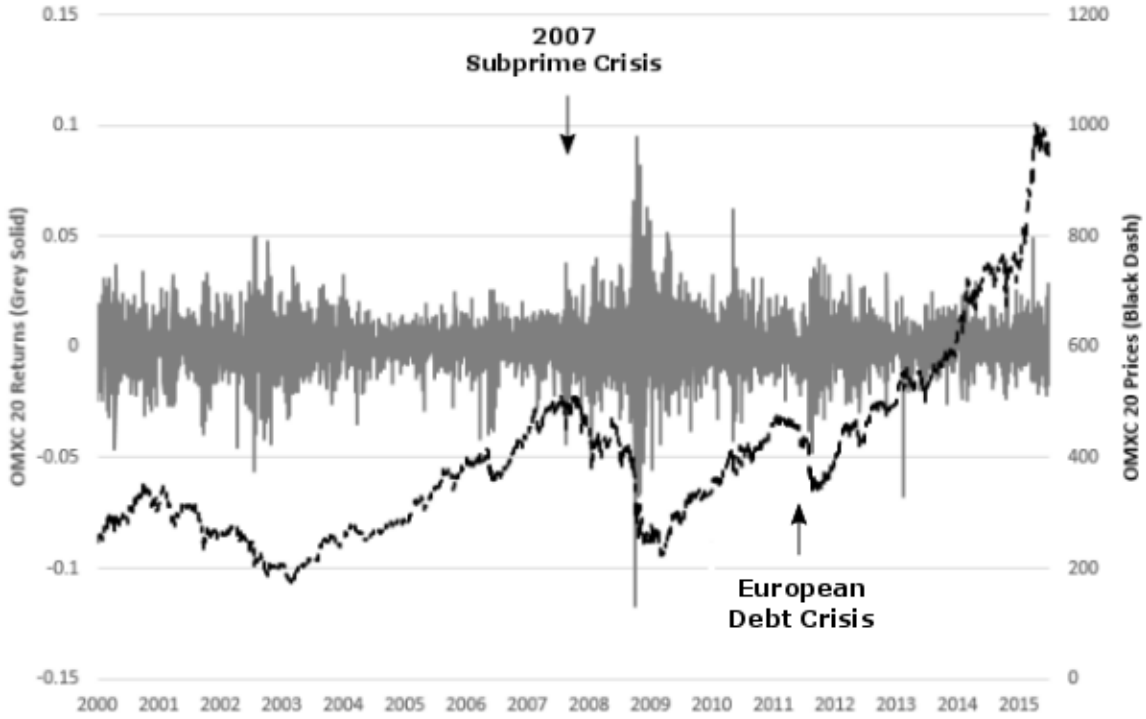


Figure 4.3: **OMXC 20 price and return, 2000-2015**

and are listed both on an American stock exchange and a Non-US stock exchange. 19 are S&P 500 components and 14 are DJIA components.

Danish Stock Market Data

The Danish economy, with a population of 5.6 million, was ranked 37^{th} by nominal GDP in the world in 2017 by the International Monetary Fund (IMF)[39] and 10^{th} in the Organisation for Economic Co-operation and Development (OECD)[68] in 2012. The stock exchange turnover was 172.5 billion Euro from Q4 2006 to Q3 2007. The key index is OMXC 20. The data collected covers the period from 01/01/2000 to 30/06/2015 (Figure 4.3).

Chinese Stock Market Data

China has been developing rapidly over the last three decades. With a population of 1.3 billion, the Chinese economy is the 2^{nd} largest in the world by nominal GDP and the largest by purchas-

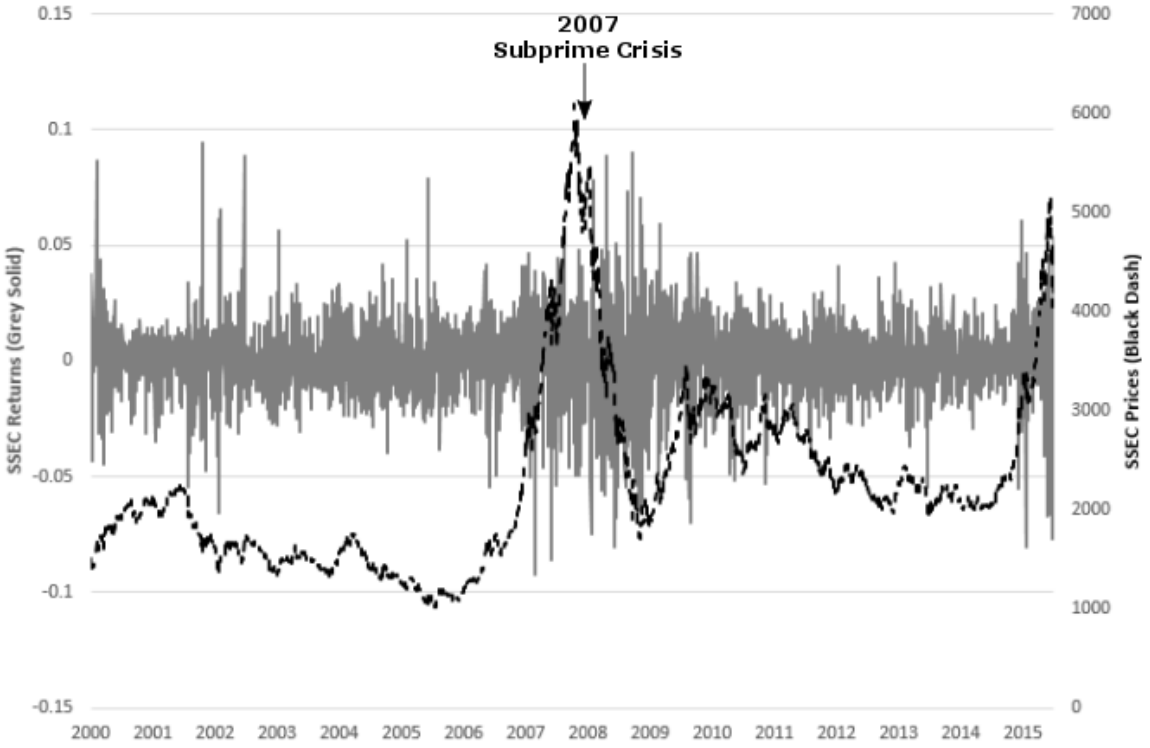


Figure 4.4: **SHCOMP price and return, 2000-2015**

ing power parity according to the IMF in 2014[39]. The Shanghai Stock Exchange Composite (SHCOMP) is selected and this market shows constant volatility during the last 15 years. The SHCOMP index is a traded stock market index[2] based on the returns of the top 50 listed companies in China by market cap. The daily time series of SHCOMP index prices is used over the same period as the Danish market data (Figure 4.4).

4.2.2 Qualitative Data - The Text Cline

US Qualitative Data

Tetlock (2007)[92] used the WSJ's daily "Abreast of the Market" (AofM) column over the 16-year period 1984-1999 as a source for investor sentiment. The same data is collected and also

[2]Like listed companies, the stock market indices can be traded as well such as DJIA, SP500, SHCOMP, etc.

Table 4.1: **Summary of benchmark corpus** – WSJ's "*Abreast of the Market*" column from 02/01/1984 to 17/09/1999 and its extension from 01/01/2000 to 31/12/2007 and business and economic news reports in WSJ and NYT from 01/01/2000 to 30/06/2015.

News source	**WSJ**		**NYT**	
News type	**AofM column**		**News reports**	
Period	**1984-1999**	**2000-2007**	**2000-2015**	**2000-2015**
Coverage (days)	4,035	2,115	4,562	5,644
Number of articles	4,039	2,121	28,838	45,008
Total number of tokens	4,701,221	2,138,913	22,294,588	44,237,531
Number of tokens per year	293,826	267,364	1,393,412	2,764,421
Number of tokens per day	1,165	1,011	4,887	7,838

- US Reportage
  - US Newspapers:
    - Major US Newspapers – *New York Times* (NYT), *Wall Street Journal* (WSJ) and etc.
    - *McClatchy-Tribune Business News*
  - US Newswires:
    - *Associated Press Financial Wire*
    - *Business Wire*
    - *PR Newswire*
    - *Standard & Poor's Daily News*
- US Editorials and Commentaries
  - Editorial & Commentary: Major US newspapers
  - Opinion: *Abreast of the Market* (WSJ) and others

Figure 4.5: **Corpora Sources**

extended to 2007 [3] to validate the techniques as the benchmark analysis for this research (see Table 4.1). Economic and business news reports are then collected from the WSJ and NYT for the period from 2000 to 2015.

US Firm Level Qualitative Data

Moving from the news articles about market indices, it was later focused on the firm-level textual sentiment of 23 large firms from the Fortune 500 list. We used two major online information

[3]The Abreast of the Market column changed the publication frequency from daily to weekly after 2007.

Table 4.2: **Summary of corpora**

	Denmark	China	23 Firms			
From	11/02/2002	01/01/2000	01/01/2000			
To	30/06/2015	30/06/2015	30/06/2015			
			US News	US Wire	US Edit&Comm	Total
Number of days	3,610	5,078	5,660	5,520	4,431	5,660
Number of articles	18,953	27,194	337,729	515,546	9,590	862,865
Number of tokens in total	2,724,171	17,366,369	217,378,508	400,220,983	6,485,714	624,085,205
Number of tokens per year	194,584	1,085,398	14,024,420	25,820,709	418,433	41,424,852
Number of tokens per day	755	3,420	38,406	72,491	1,463	145,625

sources – LexisNexis[4] and ProQuest[5] – that provide historical and up-to-the-minute news content targeted on individual companies and on industry sectors. The providers can cluster news on a historical basis, and on user-selected keywords and on relevant topics. There are three separate sources (News Cline) containing four sets of corpora that are used to collect news articles for each of these companies respectively. These sources are in Figure 4.5.

The corpus covers the period from Jan 2000 to Jun 2015, comprising over 0.8 million news reports and 624 million terms. This corpus is almost 46 times (or 32 times) as big as the Danish corpus (the Chinese corpus) (see Table 4.2).

During the period 2000 to 2015, all the return series showed a number of periods of clustering, particularly during the famous financial crisis. According to NBER announcements (NBER, 2010)[69], the period chosen has two troughs and two peaks in the international business cycle – peaks in March 2001 and December 2007 and troughs in November 2001 and June 2009. First the statistical (empirical) findings (also known as the stylised facts) are analysed for each time series to investigate whether the markets are efficient.

For the company's corpus that has been collected, each corpus is different in size (see Table 4.3). The largest corpus is Microsoft's, which contains over 173 thousand articles and close to 130 million tokens. The smallest corpus is Total's 5,000 articles and 4 million tokens. Data on the number of "mentions per article" is obtained by counting the frequency of company name appearances. On average, each company has been mentioned more than three times in each article. However, company mentions do not appear in every single news article (as part of the searching criteria was to collect news articles under the correct topics but a certain keyword doesn't have to show).

There may be many reasons that the relationship between sentiment and price movement - negative sentiment vs lower return, correlated indirectly - not holding at the firm level of the market. First, as it is not easy to obtain a description of a firm exclusively in a newspaper text

[4]This provider indexes its reportage and editorials for business professionals and is designed for general news provision, company research and due diligence.

[5]This is an information provider that targets libraries and research organisations: it is a key partner for content holders of all types, preserving and enabling access to their rich and varied information.

Table 4.3: **Summary of each firm's corpus**

	Industry	**Number of Articles**	**Number of Tokens**	**Mentions Per Article**
IBM	IT	34,801	25,514,725	5.89
Wal-Mart	RE	48,304	30,777,104	5.48
Toyota	AU	20,226	13,187,342	5.12
Cisco	IT	34,995	25,462,071	5.01
Apple	IT	57,663	38,166,045	4.44
Verizon	IT	82,382	59,315,624	4.28
Microsoft	IT	173,757	127,403,166	4.00
AT&T	IT	59,189	45,020,086	3.98
Merck	PH	24,087	21,469,385	3.95
Chevron	OI	13,379	9,728,932	3.76
Boeing	AE	51,762	35,601,713	3.73
Intel	IT	65,008	47,520,377	3.59
HP	IT	68,508	50,494,450	3.49
HomeDepot	RE	17,721	11,741,127	3.46
Conoco	OI	14,343	11,489,829	2.77
Pfizer	PH	40,252	31,286,611	2.62
Johnson&Johnson	MA	25,872	18,575,764	1.56
Exxon	OI	23,661	16,118,207	1.04
Ford	AU	48,785	34,183,082	1.00
Shell	OI	17,275	12,227,512	0.99
GE	EE	64,866	46,707,054	0.82
BP	OI	18,291	13,485,306	0.72
Total	OI	5,304	4,222,483	0.37
Average		43,932	31,726,000	3.13

Table 4.4: **Percentages of companies mentioned in each of the individual company's corpus** Note: the percentage of company mentions denotes the number of articles that the company name has been mentioned compare to the total number of articles.

	Industry	**Company Mentions**	**Overlapping with Other Companies**			
Toyota	AU	91%	3%	Ford	1%	Microsoft
Wal-Mart	RE	91%	2%	HomeDepot	1%	Apple
Boeing	AE	89%	2%	Microsoft	1%	IBM
HomeDepot	RE	82%	8%	Wal-Mart	1%	Microsoft
Merck	PH	81%	9%	Pfizer	2%	Johnson&Johnson
IBM	IT	78%	6%	Microsoft	4%	HP
Chevron	OI	77%	5%	Exxon	5%	Conoco
Verizon	IT	77%	11%	AT&T	3%	Apple
Cisco	IT	75%	5%	Microsoft	4%	IBM
Microsoft	IT	74%	5%	Apple	4%	HP
Pfizer	PH	73%	10%	Merck	4%	Johnson&Johnson
Apple	IT	73%	10%	Microsoft	4%	AT&T
AT&T	IT	72%	12%	Verizon	4%	Apple
Conoco	OI	68%	11%	Chevron	7%	Exxon
HP	IT	64%	9%	Microsoft	8%	IBM
Intel	IT	62%	10%	Microsoft	7%	HP
Johnson&Johnson	MA	59%	9%	Pfizer	8%	Merck
Ford	AU	50%	24%	Toyota	5%	Microsoft
Shell	OI	49%	13%	Chevron	10%	Exxon
GE	EE	47%	11%	Boeing	6%	Microsoft
BP	OI	43%	13%	Chevron	11%	Exxon
Exxon	OI	38%	14%	Chevron	10%	Conoco
Total	OI	24%	16%	Chevron	15%	Exxon
Average		67%	10%		5%	

- usually there is news also about firms in the same category: you cannot have an oil company mentioned in isolation very often; it is also true for some IT firms (see Table 4.4). The percentage of the company mentions shows the number of mentions of the specific company name occurs in each of the company's corpus and the percentage of overlapping shows mentions of other companies rather than the target company in the corpus. The results show that industries like retail and IT has a higher percentage of company mentions while oil and manufacturing industries, in general, has a lower percentage of company mentions. In the meantime, the corpus of oil industry companies tends to have higher overlap in mentions of the company names in the same industry.

Danish Qualitative Data

Danish economic and financial news published in 36 newspapers (originally in Danish and also available in English translation) is then collected on a daily basis. The news articles are downloaded from the business news provider "*LexisNexis News and Business*". We searched for maximum availability of news articles starting from 11 February 2002 and ending on 30 June 2015. We used "*Denmark*" or "*Danish*" as keywords and "*Banking & Finance*" or "*Economy & Economic Indicators*" as industry and subject in the search criteria. All the news in English are then collected from "*Danish Business Digest*"[6] and "*Esmerk Denmark News*"[7]. After combining news articles from different sources, the Danish economic news corpus comprised nearly 19,000 news articles which include over 2.7 million tokens[8] (Table 4.2).

Chinese Qualitative Data

In China, the more widely used source of business information is the *Xinhua news agency* (Xinhua), which is government owned. Xinhua's output is translated into English. This translation is available daily from the news aggregator LexisNexis with frequent revisions throughout the day.

[6]Danish Business Digest is a daily abstracting service in English, which covers Denmark and provides corporate, industry and economic news.

[7]Esmerk Denmark News provides English-language summaries on key business issues abstracted from local language sources (including 96 different news agencies: Børsen, Jyllands-Posten and Politiken etc).

[8]A **token** is an individual occurrence of a linguistic unit in speech or writing.

It is assumed that Xinhua's output is an authentic account of business and finance in China. We downloaded news articles again from "LexisNexis News and Business". "LexisNexis News and Business" annotates each news item with a large set of keywords that fall into categories like politics, technology, and business and finance. Within the business and finance category, we searched for a cluster of terms, internal to LexisNexis but under two broad headings "Banking & Finance" and "Economy & Economic Indicators". Moreover, we set search criteria such that tokens "China" or "Chinese" have to occur 3 or more times in each news item in order to make sure that output news articles are primarily dealing with a Chinese event. The result is a corpus of texts from Jan 2000 to Jun 2015 which contains over 27,000 articles with around 17 million terms (or tokens) (Table 4.2).

In summary, the news cline is 'complete' for the US market at the aggregate and firm levels. For the Danish and Chinese markets, we have to rely on the English translations that are digitally available. It is believed that a good coverage of qualitative data (see Table 4.5) has been established.

Table 4.5: **News Cline of All Markets**

Market	News Report	Editorial/Commentary	Newswire
US DJIA Index	√	√	√
US Firms	√	√	√
Danish OMXC Index	√		
Chinese SHCOMP Index			√

4.3 Market Inefficiency: Stylised Facts

4.3.1 Market Level Inefficiency

As one of the largest industrial indices in the globe, based on the 30 largest publicly owned companies in the US, the DJIA index price has climbed nearly ten times from $1,199 to $10,804 during the fifteen-year period of 1984 to 1999 (the same period that Tetlock (2007)[92] analysed in his paper). The daily return series r_t for the price is the logarithm of the ratio of the price

today divided by price yesterday. The summary statistics (in Table 4.6) show that the mean value of DJIA returns during that period was 5.83 basis points. Extreme negative skewness and positive kurtosis indicate that the distribution of the returns series is not normal. Three types of annual returns suggest that on average the index provides a return rate of around 15% per annum. The series has a significant z-score value that rejects the null hypothesis of zero expected return.

Table 4.6: **Summary statistics for times series of returns for DJIA. N=3691:** The observations start on Jan 01, 1984 and end on Sep 17, 1999. G%, A% and A*% denote constant annual return, arithmetic annual return and consecutive annual return respectively. The last columns has the z-score for each series.

Series	$10^4\bar{r}$	$10^2 sd$	**skewness**	**kurtosis**	**G%**	**A%**	**A*%**	**z**
DJIA	5.83	1.06	-4.18	100.06	15.7	14.2	17.3	3.35

It is clear that there is a climbing trend of the DJIA price series and a number of periods where extreme returns are clustered (see Figure 4.1). Clustering suggests that conventional wisdom, that for every up-tick there is a down-tick, does not quite work; when the market starts showing instability in this manner one must look for other causal explanations. Large stock indices probably move on their own. However, some small indices may be highly dependent on some of the major indices. Further to the analysis of the major indices in the world, we also looked at some representative indices for EU and emerging markets.

In order to discover the possible intersection between indices of different sizes, data is used from the DJIA, S&P 500, OMXC 20 and SHCOMP indices covering the same period from Jan 2000 to Jun 2015. Similar to the DJIA, the S&P 500 is based on the market capitalisations of 500 large companies in America. OMXC 20 is a weighted index comprising 20 most-traded stocks in Denmark, and SHCOMP is a weighted index based on all stocks (A shares and B shares) that are traded on the Shanghai Stock Exchange.

The index price time series contain information over the period 01/01/2000 through 30/06/2015. Ideally, the return is expected to follow a random walk – every up-tick followed by a tick-down. The mean values of the return series for DJIA, S&P 500, OMXC 20, and SHCOMP are close

to zero (1.10×10^{-4},0.87×10^{-4},3.37×10^{-4}, and 3.04×10^{-4})[9], however the distribution of the return series is not normal, as their excess kurtosis are greater than zero and the series are skewed negatively (Table 4.7). G%, A% and A*% show that American indices have smaller annual returns than OMXC and SHCOMP.

4.3.2 A Note on Market Volatility

The results from the GARCH(1,1) tests (Equation 3.4) for all five market indices can be explained by the coefficients $\alpha^{'}$ - weight on past returns and $\beta^{'}$ - weight on past variances. The econometrician must estimate the constants ω,$\alpha^{'}$, $\beta^{'}$; updating simply requires knowing the previous forecast h and residual. The weights are $(1 - \alpha^{'} - \beta^{'}$, $\beta^{'}$,$\alpha^{'})$ and the long run average variance is $\sqrt{\omega/(1 - \alpha^{'} - \beta^{'})}$. It should be noted that this only works if $\alpha^{'} + \beta^{'} < 1$, and only really makes sense if the weights are positive requiring $\alpha^{'} > 0$, $\beta^{'} > 0$, $\omega > 0$ .It indicated that $\alpha^{'}$ is around 0.1 and $\beta^{'}$ close to 0.9 (Table 4.8). The persistence $(\alpha^{'} + \beta^{'})$ is close to, and less than, unity in all cases as return series of these indices revert to the mean value. A higher $\alpha^{'}$ (or $\beta^{'}$) normally indicates the variance of returns series is dependent on its past squared returns (or conditionally dependent on its past variance). SHCOMP has the highest $\beta^{'}$ of the five indices, with relatively lower $\alpha^{'}$ and (less significant ω). OMXC 20 has the largest $\alpha^{'}$. Although the persistence of S&P 500 and SHCOMP are same, the $\alpha^{'}$ of S&P 500 (0.093) is higher than it is for SHCOMP (0.076), indicating that S&P 500 is perhaps a more efficient market than the SHCOMP.

Volatility clustering is perhaps one manifestation of irrational behaviour. As this behaviour is motivated by sentiment, the lack of auto-correlation and the presence of volatility clusters provides prime facie evidence of the key role played by sentiment and justifies this investigation. Note that as it is decided to use autoregressive techniques, which rely on volatility clustering and constant variance, the claims relating to the impact of sentiment may be weakened.

Other statistical tests like the Pearson's correlation coefficients measure the degree of linear association between markets. The results show that the correlation between DJIA and S&P 500

[9]Results of z tests show that the null hypotheses of zero mean values are all accepted.

Table 4.7: **Summary statistics for times series of returns for selected indices and firms.** The observations start in Jan 2000 and end in Jun 2015. G%, A% and A*% denote constant annual return, arithmetic annual return and consecutive annual return respectively. The last columns are the z-scores for each series to access the null hypothesis that the expected return is zero.

	Count	10^4 r	10^2 sd	Skewness	Kurtosis	G%	A%	A*%	z
				Indices					
DJIA	3897	1.10	1.19	-0.06	8.08	2.79	2.75	4.63	0.58
OMXC	3873	3.37	1.30	-0.23	5.45	8.80	8.43	11.13	1.61
SHCOMP	3750	3.04	1.60	-0.17	4.31	7.64	7.36	11.04	1.16
Max	3897	3.37	1.60	-0.06	8.08	8.80	8.43	11.13	1.61
Min	3750	1.10	1.19	-0.23	4.31	2.79	2.75	4.63	0.58
				Firms					
Apple	3897	9.22	2.88	-4.33	109.47	26.09	23.18	39.96	2.00
AT&T	3897	1.08	1.71	0.10	6.94	2.74	2.71	6.60	0.39
Boeing	3897	3.91	1.96	-0.26	5.71	10.34	9.84	15.77	1.25
BP	3897	0.56	1.81	-0.42	10.30	1.43	1.42	5.70	0.19
Chevron	3897	3.38	1.63	0.06	11.85	8.87	8.50	12.57	1.29
Cisco	3897	-1.44	2.63	0.16	8.06	-3.56	-3.63	5.23	-0.34
Conoco	3897	4.42	1.84	-0.42	6.68	11.76	11.12	16.61	1.50
Exxon	3897	2.79	1.58	0.02	10.75	7.28	7.02	10.69	1.11
Ford	3897	-0.82	2.83	0.00	13.91	-2.04	-2.06	8.32	-0.18
GE	3897	-0.49	1.99	0.06	8.36	-1.21	-1.22	3.84	-0.15
HomeDepot	3897	1.98	2.10	-0.96	21.56	5.12	4.99	11.08	0.59
HP	3897	-0.44	2.48	-0.26	7.95	-1.11	-1.11	6.85	-0.11
IBM	3897	1.61	1.69	-0.08	7.83	4.13	4.04	7.95	0.59
Intel	3897	0.03	2.49	-0.46	8.42	0.08	0.08	8.21	0.01
Johnson	3897	2.88	1.24	-0.50	16.01	7.51	7.25	9.62	1.45
Merck	3897	1.15	1.82	-1.53	28.83	2.93	2.89	7.30	0.39
Microsoft	3897	0.17	2.01	-0.13	9.47	0.44	0.44	5.69	0.05
Pfizer	3897	1.39	1.65	-0.26	5.25	3.54	3.48	7.13	0.53
Shell	3897	1.79	1.76	-0.09	7.88	4.61	4.50	8.76	0.64
Total	3897	2.66	1.81	-0.13	5.30	6.92	6.69	11.41	0.92
Toyota	3897	1.05	1.76	-0.12	7.23	2.69	2.65	6.76	0.37
Verizon	3897	1.34	1.64	0.17	6.53	3.43	3.37	6.99	0.51
Wal-Mart	3897	0.71	1.54	0.13	5.61	1.81	1.79	4.90	0.29
Max	3897	9.22	2.88	0.17	109.47	26.09	23.18	39.96	2.00
Min	3897	-1.44	1.24	-4.33	5.25	-3.56	-3.63	3.84	-0.34

Table 4.8: **Estimation of a GARCH(1,1) model for daily log-returns.**

Series	**DJIA**	**SP500**	**OMXC**	**SHCOMP**
ω	0.000002	0.000002	0.000005	0.000003
	(0.0002)	(0.083)	(0.001)	(0.023)
α'	0.099	0.093	0.105	0.076
	(0.000)	(0.000)	(0.000)	(0.000)
β'	0.889	0.896	0.865	0.913
	(0.000)	(0.000)	(0.000)	(0.000)
$\alpha' + \beta'$	0.988	0.989	0.970	0.989
log-lik.	12609.63	11964.17	11113.24	10309.24

is in excess of 50% (Table 4.9). American indices are around 40% correlated with OMXC 20 and 4% correlated with SHCOMP. Commodity and Purchasing Managers' Indicator (PMI) indices are positively correlated with market indices but the Dollar (DI) and Volatility (VIX) indices are correlated negatively. However, if we take a look at the lagged markets correlations (Table 2.4) again, it is reasonable to believe that due to time zone differences, OMXC and SHCOMP are in fact correlated with other markets delayed by one day. Therefore, the conclusion is that all the market indices are correlated and they are not suitable to be put into regression models as independent variables as there is information overlap between markets and they are not actually 'independent'.

4.3.3 Firm Level Inefficiency

Moving from stylised facts for index return series, now it is focused on properties of some firms' return series. The selected 23 companies are all from Fortune 500 Global and Table 4.7 shows various moments including the mean, standard deviation, skewness, kurtosis, and z-statistics. The mean values of the return series for all the firm-level stock returns are slightly positive, with high standard deviations and non-normal distributions (Table 4.7). In general, the firm stock returns' kurtosis is higher than indices'. Among these companies, only the mean value of Apple's return series is significantly different from zero in z-statistic tests. Apple has the highest mean value, standard deviation, kurtosis, constant, arithmetic and consecutive annual returns, but the

Table 4.9: **Correlations between times series of returns for four indices, Euronext Rogers International Commodity Index (RICI), Federal Reserve Trade Weighted U.S. Dollar Index (DI), PMI (Monthly) Index and 23 selected firms.** The observations start in Jan 2000 and end in Jun 2015.

	DJIA	*SP500*	*OMXC*	*SHCOMP*	*VIX*	*DI*	*PMI*	*RICI*
Indices								
SP500	97%							
OMXC	39%	40%						
SHCOMP	4%	4%	12%					
Indicators								
VIX	-72%	-75%	-33%	-5%				
DI	-10%	-12%	-14%	-8%	8%			
PMI	21%	22%	18%	23%	-5%	-14%		
RICI	24%	28%	28%	10%	-22%	-39%	30%	
Firms								
Apple	45%	50%	15%	3%	-37%	-3%	20%	10%
AT&T	59%	58%	24%	2%	-40%	-7%	8%	15%
Boeing	65%	61%	30%	3%	-48%	-8%	28%	20%
BP	58%	59%	32%	6%	-45%	-23%	24%	41%
Chevron	66%	66%	31%	7%	-51%	-20%	20%	45%
Cisco	58%	64%	23%	2%	-45%	-5%	22%	15%
Conoco	60%	62%	30%	7%	-48%	-20%	26%	47%
Exxon	68%	67%	26%	6%	-53%	-15%	14%	39%
Ford	54%	53%	27%	4%	-39%	-5%	21%	12%
GE	75%	75%	31%	0%	-54%	-10%	15%	17%
HomeDepot	65%	63%	24%	0%	-46%	-2%	13%	7%
HP	56%	58%	25%	3%	-42%	-5%	17%	12%
IBM	66%	64%	27%	3%	-46%	-7%	29%	14%
Intel	62%	65%	23%	2%	-46%	-3%	34%	13%
Johnson	57%	53%	16%	0%	-40%	-3%	25%	10%
Merck	52%	50%	19%	4%	-37%	-4%	12%	12%
Microsoft	63%	66%	25%	3%	-48%	-4%	21%	15%
Pfizer	58%	57%	22%	1%	-43%	-4%	4%	12%
Shell	64%	64%	34%	6%	-48%	-27%	20%	44%
Total	65%	66%	38%	8%	-54%	-33%	22%	46%
Toyota	55%	56%	29%	6%	-43%	-12%	14%	18%
Verizon	58%	57%	21%	2%	-41%	-6%	10%	11%
Wal-Mart	57%	54%	14%	1%	-40%	1%	8%	-1%
Max(Firm)	75%	75%	38%	8%	-37%	1%	34%	47%
Min(Firm)	45%	50%	14%	0%	-54%	-33%	4%	-1%

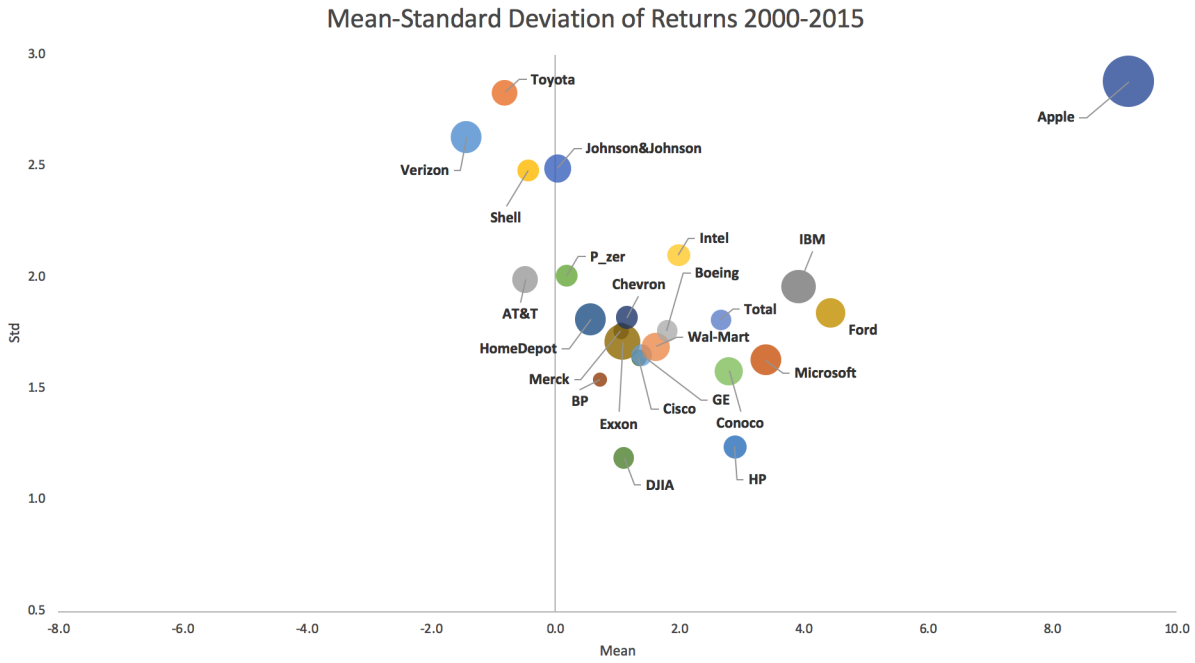


Figure 4.6: **23 companies' returns mean vs standard deviation, 2000-2015**

lowest skewness. In other words, Apple's stock prices have the most extreme values among these companies.

Comparing the mean and standard deviation of each company's returns, and draw this relationship on the graph (Figure 4.6, 4.7 & 4.8), one can observe that the company's performance compared with DJIA (the circle size represents the value of the chi-square test of the sentiment impact in this period). During the entire period, the closest company to the Dow Jones index is HP (low return low risk) and the most distant is Apple (high return high risk). Most IT and crude oil companies are clustered between low returns (< 0.0002) and medium risks ($1.5 <> 2.5$). When the 2000-2007 period is observed, the distribution of companies is relatively better dispersed. During this period, Apple is basically in the same position as DJIA, while crude oil companies are relatively moved to high return areas ($0.0003 < \& < 0.0008$). When switching to the 2008-2015 period, most crude oil companies are in the same return range (with DJIA), but the risk is relatively high, and most IT companies have relatively high returns.

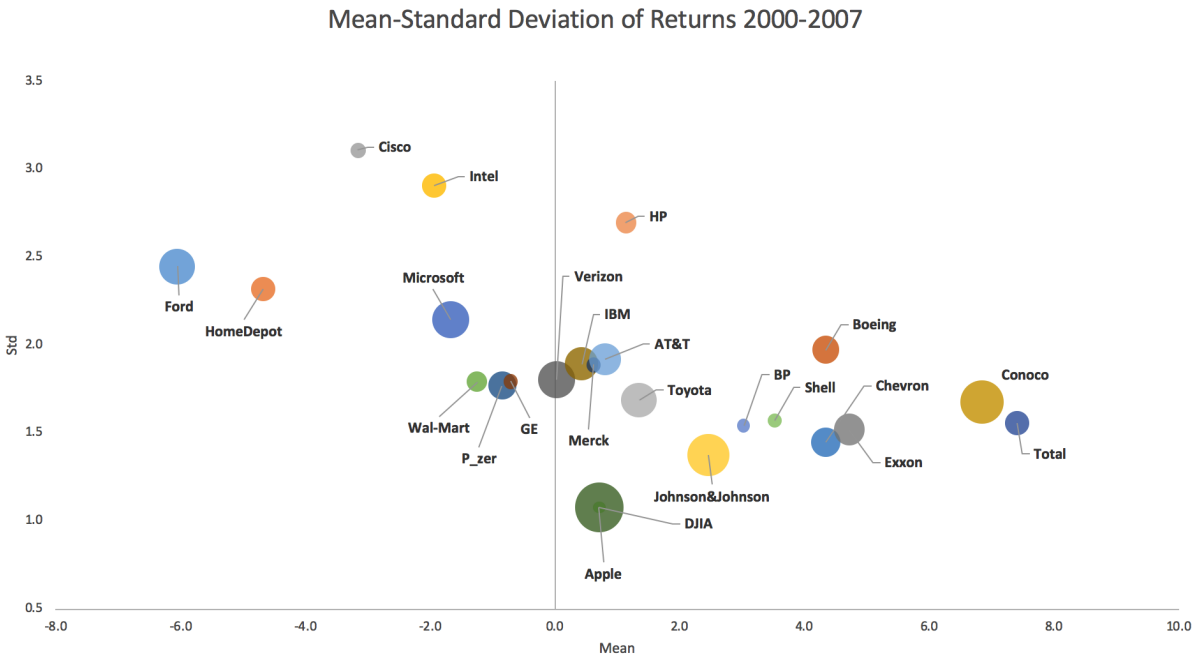


Figure 4.7: **23 companies' returns mean vs standard deviation, 2000-2007**

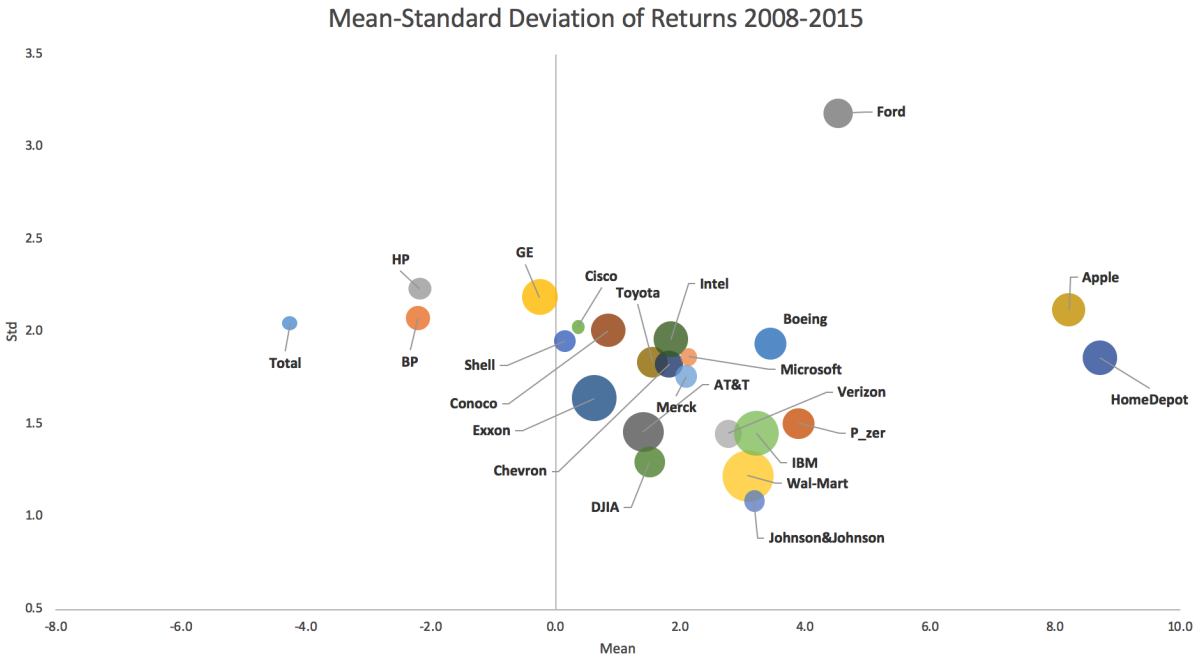


Figure 4.8: **23 companies' returns mean vs standard deviation, 2008-2015**

The performance of each company varies even if some companies are in the same industry or sector. Looking at the Pearson correlation coefficients, all returns are positively correlated with market indices, PMI indicator and commodity index; and negatively correlated with the Dollar index and the VIX volatility index. Correlation coefficients between firms and DJIA are all ranged between 45% and 75%. The highest figure is for General Electric and the lowest is for Apple. As a comparison, the correlation between firms and commodity market are distributed between -1% and 47%. All oil companies from the list have a correlation around or higher than 40%. Only Wal-Mart is negatively correlated (-1%). The Dollar index is negatively correlated with most firms, though oil and manufacturing companies tend to have lower correlations, except the outlier Wal-Mart. The lowest correlation (-33%) is between the Dollar index and Total and the highest (1%) is between the index and Wal-Mart. The purchasing power index, PMI, has higher correlations with companies who are selling products rather than oil companies or service providers.

In conclusion, the fact that individual return series are not auto-correlated. It is important to note that the 5 day-lag of return series values are independent of each other and mean reversing over time. By analysing the stylised facts of return series, the results suggest that although prices might have caused the rises (or falls) by the peaks (or troughs) of themselves, however, the returns of prices has not causality among different day lags. This corresponds to one of the stylised facts of the returns (Taylor,2011)[90] that there is almost no correlation between returns for different days and avoids any confounding effects in the sentiment analysis caused by the autocorrelation of the prices. A correlation between different indices indicates that return series are not entirely independent then because there are interceptions between each market and these relationships are tested later in the VAR analysis.

4.4 Case Study I – Benchmark: DJIA

First of all, it is attempted to replicate and extend Tetlock's (2007)[92] results following the same methodology with the same financial and textual data. Essentially, we are looking at

the relationship between DJIA returns and their past historical values on a daily basis over a 15-year period (Jan 1984 Sep 1999) together with the investor sentiment (proxy) during the same period. The analysis then is extended from 1984 until 2007 (after 2007 the Abreast of the Market column changed from daily release to weekly). We also extend the news sources using business and economic news from WSJ and NYT over a 15-year period from 2000 to 2015. A VAR model is conducted (Equation 3.5) to test the impact of their historical measures. The assumption of this model is that the expectation of regression residuals is independently and identically distributed (i.i.d). The regression of the return variable was conducted using a number of endogenous variables, including five lags of DJIA returns, traded volumes, and negative sentiment[10], and exogenous variables, including dummy variables of day-of-the-week, month-of-the-year, US holidays and the 1987 financial crash effects as well as five lags of conditional volatility measure[11]. The measure of traded volume is the detrended log volume. Newey and West (1987)[67] robust standard errors are used to reduce the Heteroskedasticity of residuals.

The historical impact of the WSJ's AofM negative sentiment on DJIA current return is tested using the estimated model as follows:

$$R_t = \alpha + L_5\beta R_t + L_5\gamma lnV_t + L_5\delta Sent_t + Exog_t + \varepsilon_t \tag{4.1}$$

The coefficients (α, β, γ, and δ) help in quantifying the impact of historic values of the dependent variable and that of the control variables like sentiment and stock market movements elsewhere. The regression equations determine the values of these coefficients. The values of these coefficients, together with the values of their statistical significance, are an indication of the impact of each of the variables on the right-hand side of equations.

[10]The sentiment variable is de-meaned in the calculations however the results are similar if the sentiment variable is not de-meaned during tests.

[11]Volatility measures the price variability, basically the standard deviation of returns, over a time period. We look at the properties of the return time series and volatility measures assist in explaining part of the returns' abnormal movements. There are many ways of quantifying the volatility in the past research studies. Most popular ones are absolute values of return series, conditional standard deviations and VIX index. Each of them has its pros and cons. We use the conditional variance series computed from GARCH(1,1) model as the volatility measure in this research. Compare to unconditional variance,

Table 4.10: **Benchmark Index: impact of negative sentiment extracted from WSJ – Abreast of the Market column, and WSJ and NYT – business & economic news on DJIA index returns:** dependent variables are DJIA returns and all coefficients are in **basis points**.

Study	Replica of Tetlock's paper						Our extended studies					
Corpus	WSJ-AofM		WSJ-AofM		WSJ-AofM		WSJ-AofM		WSJ-News		NYT-News	
Period	1984-1999		1984-1991		1992-1999		2000-2007		2000-2015		2000-2015	
Observations	3709		1714		1994		1992		3682		3858	
Neg_t												
$L_1\delta$	-4.8	***	-4.3		-7.4	***	0.1		2.3		-2.0	
$L_2\delta$	2.2		1.0		3.9		-3.0		-0.5		-0.7	
$L_3\delta$	-1.0		-2.7		0.2		1.3		2.2		1.3	
$L_4\delta$	4.9	***	7.1	**	4.1		-2.7		2.8		2.3	
$L_5\delta$	2.5		0.6		4.3		4.4	**	-4.8	**	-0.5	
$\chi^2(5)$[Joint]	20.20	***	7.07		15.60	***	8.10		5.85		2.55	
p-value	(0.001)		(0.226)		(0.008)		(0.151)		(0.321)		(0.769)	
Sum of 2 to 5	8.60	***	5.98		12.42	***	-		-		2.34	
$\chi^2(1)$[Reversal]	8.59	***	1.11		7.82	***	-		-		0.62	
p-value	(0.003)		(0.291)		(0.005)		-		-		(0.430)	

First of all, the values of DJIA returns (1984-1999, N=3691 and 2000-2015, N=3882) is regressed against its previous 5 days values in the first instance to see the impact of the historic values of DJIA on its present value. None of the day lags has a statistically significant contribution level ($p < 0.1$) confirming the property of stock returns that there is no correlation between returns for different days.

Then, we looked at the impact of negative sentiment on DJIA returns. The first model to be computed is based on Tetlock's 1984-1999 period data. The negative impact of 1^{st} day lag negative sentiment is 4.8 basis points and is statistically significant at the 0.01 level; this is fully recovered within the 2^{nd} to 5^{th} days lag (with fourth-day lag statistically significant at 0.01 level). This period is then split into two: 1984-1991 and 1992-1999 as Tetlock did in his paper to test the difference between normal and bull periods. The normal period (1984-1999) includes the famous Black Monday crash but the bull period doesn't have any business downturns (according to NBER business cycles). The results show that the impact of negative sentiment is more significant during the 1990s (negative 7.4 basis points 1^{st} day lag significance at 0.01 level) than during the 1980s (positive 7.1 basis points 4^{th} day lag significance at 0.05 level). The data period is further extended to the volatile 2000s (there are two recessions over the period 2000 to 2007). The results are similar to those in the 1980s; the 4.4 basis points lag five of negative sentiment is significant at 0.05 level. After the data period extension, we chose different news sources for testing the variable impact of the news cline. Negative sentiment is extracted from WSJ and NYT news reports. Clearly, negative sentiment in various news sources has a different impact on the DJIA index. (see Table 4.10)

The results of the experiments correspond to Tetlock's results. So far, it is confident in implementing the mentioned methodology into further tests on the impact of negative sentiment in representative EU and Asian markets.

---

conditional variance is not finite. The value at time t depends on all previous value from time 1 to t-1. All the values are adjusted by time domain and not fixed over the observation period.

4.5 Case Study II – Danish Economy and the Quantitative and Qualitative Sentiment Proxies

Essentially, it is essential to know what is the time domain development of OMXC 20 and S&P 500 in terms of their past historical values on a daily basis over a 13 year period (Feb 2002-Jun 2015) together with the evolution of sentiment variable.

A multivariate time-series model will be utilised in this study. We treated all the variables as endogenous and regressed each of the variables on past lags of these variables (Equation 4.1). The models are fitted using ordinary least squares estimation techniques. Following methodological conventions, we tested hypotheses by assessing the joint statistical significance of the coefficients on single variables and using chi-square tests.

Following the defined 5^{th} order Vector Autoregressive 4-step model (Table 3.1) with the error term ϵ_i, we tested the hypotheses from steps I and III (as a result of the indices inter-correlations, an external index cannot be used in the regression model) and output the results (Table 4.11).

Table 4.11: **Impact of negative sentiment on OMXC 20 returns:** *, ** and *** denote values of coefficients' (α, β, γ, and δ) statistical significance at 0.1, 0.05 and 0.01 levels respectively. All coefficients are in **basis points**.

		Dependent: r_t^{OMXC}		
Tests		**Step I**	**Step III**	
βr_t^{OMXC}	L_1	319	318	
	L_2	-264	-259	
	L_3	-402	-402	
	L_4	378	379	
	L_5	-437	-431	
γNeg_t^{Danish}	L_1		-6.0	**
	L_2		0.9	
	L_3		0.6	
	L_4		-0.7	
	L_5		4.4	*
Neg_t^{Danish}	χ^2**(5)[joint]**		5.13	

The values of OMXC 20 returns (2002-2015, N=3143) is regressed against its previous 5 days values in the first instance to see the impact of the historic values of OMXC 20 on its present value.

The 1^{st} lagged return values have a positive contribution followed by negative contributions from the 2^{nd}, 3^{rd} and 5^{th} day lags and a positive contribution from the 4^{th} day lag. The overall contribution is negative. However, none of the day lags has a statistically significant contribution level ($p < 0.1$), confirming the property of stock returns that there is no correlation between returns for different days. The results are shown in Table 4.11 – "Step I".

When the negative sentiment is added to the first step, the impact of the past values of OMXC 20 returns on its present value is slightly smaller in magnitude. However, the negative impact of 1^{st} day lag negative sentiment is 6.0 basis points and statistically significant at the 0.05 level ("Step III" in Table 4.11); it is nearly recovered within the 2^{nd} to 5^{th} days lag with the 5^{th} day lag having a statistically significant impact of 4.4 basis points at the 0.1 significance level.

Compared to the impact of negative sentiment on the US market, the impact on the Danish market is similar in magnitude but less significant. Additionally, the negative effect significantly recovered on the 4^{th} day in the US market, however, the recovery happened on the 5^{th} day (slower) in the Danish market.

4.6 Case Study III – Chinese Economy and the Computational Account of Investor Sentiment

In the Chinese market, we looked at the relationship between S&P500 and SHCOMP returns and their past historical values on a daily basis over a 15 year period (Jan 2000 Jun 2015) together with the investor sentiment (proxy) during the same period. The same vector autoregressive model (Equation 4.1) is applied to test the relationship between different inputs and their historical values. Again, the regression of the return variable is tested using different endogenous variables, including five lags of SHCOMP returns, traded volumes, S&P 500 returns and negative sentiments, and exogenous variables, including dummy variables of day-of-the-week and month-of-the-year effects as well as five lags of conditional volatility measure. The detrended log volume of SHCOMP index is used as the measure of traded volume.

Vector Autoregressive models are again used as the estimation technique. The errors are fitted using ordinary least squares. We also conducted chi-square tests and tested the causality relationship between endogenous variables.

The hypotheses we will test in this case study are listed in the methods section (Table 3.1). To test the hypotheses we used a 5^{th} order Vector Autoregressive model with error term ε_t. The coefficients (α, β, γ, δ, and θ) measured the sign and magnitude of the impact of past values of the dependent and independent variables.

The hypotheses on a 15-year daily data set of the endogenous and exogenous variables associated with the SHCOMP is tested. It is noted that the independence of the current value of SHCOMP return on its past values is statistically significant – this is because the series values are not auto-correlated (see Table 4.12). However, as discussed the SHCOMP and S&P 500 are inter-correlated; we, therefore, don't use the S&P 500 as an external constraint in this model (i.e. we skip Steps II and IV in the aggregation methodology).

Table 4.12: **Hypothesis tests of impact of S&P 500 and negative sentiment on SHCOMP returns and volumes**: *, ** and *** denote values of coefficients' (α, β, γ, and δ) statistical significance at 0.1, 0.05, and 0.01 levels respectively. All coefficients are in basis points.

		Dependent: r_t^{SHCOMP}	
	Tests	**Step I**	**Step III**
βr_t^{SHCOMP}	L_1	33	60
	L_2	-85	-93
	L_3	255	251
	L_4	430	435
	L_5	-172	-179
γNeg_t^{Xinhua}	L_1		-9.5 ***
	L_2		7.0 ***
	L_3		-0.5
	L_4		1.8
	L_5		-0.6
Neg_t^{Xinhua}	$\chi^2(5)[joint]$		15.5 ***

The estimation process used is to regress the values of SHCOMP returns (2000-2015, N = 3474) against its lagged values to measure the impact of the past values on its present value. Model coefficients of the hypothesis I test show no autocorrelations between lagged returns (all of these lags are not statistically significant). (Table 4.12 – "Step I").

If we add in the investor negative sentiment to the "Step I", we find that the 1^{st} day lag has 9.5 basis points negative contribution to SHCOMP returns and 2^{nd} day lag has 7 basis points reversing contribution (both of them are statistically significant at the 0.01 level)(Table 4.12 – "Step III").

The interdependence between returns and sentiment variable is tested using Granger causality tests (χ^2). Investor negative sentiment Granger causes SHCOMP returns at the 0.01 level.

Compared to the impact of negative sentiment on the US and Danish markets, the impact on the Chinese market is almost double the magnitude of the US market with similar significance (but more significant than in the Danish market). The negative effect significantly recovered on the 4^{th} day in the US market and on the 5^{th} day in the Danish market, but it significantly recovered on the 2^{nd} (much faster) in the Chinese market.

4.7 Case Study IV – Market-level Relationships between Sentiment and Return

4.7.1 Linear Relationship - VAR

First, the hypothesis sentiment has a statistically significant impact on returns is tested by incorporating 'sentiment' in the regressive model of returns using vector autoregression (VAR). In a model similar to Tetlock (2007)[92] and carried out previously by the Kelly and Ahmad (2015), Zhao and Ahmad (2015a), Zhao and Ahmad (2015b)[48][103][104], we will use the same regression (Equation 4.1) to test the hypothesis.

$L_5\delta Sent_t$ will be replaced by $L_5{S_t}^{DJIA}$, $L_5{S_t}^{OMXC}$, $L_5{S_t}^{SHCOMP}$ or $L_5{S_t}^{WTI}$ accordingly when we analysed a text corpus from different sources. Running this model allows a linear relationship to be created between the sentiment measure extracted from news and financial returns.

Table 4.13: **The negative effects of news sentiment on financial instruments**

Case Study	DJIA				OMXC [103]	SHCOMP [104]	WTI [47]
Period	1984 -1999	1984 -1991	1992 -1999	2000 -2007	2002 -2015	2000 -2015	2000 -2014
Observations	3709	1714	1994	1992	3143	3474	3553
L_nS_t							
L_1	**-4.8**	-4.3	**-7.4**	0.1	**-6.0**	**-9.5**	-2.1
L_2	2.2	1	3.9	-3	0.9	7.0	**-8.5**
L_3	-1	-2.7	0.2	1.3	0.6	-0.5	-8.5
L_4	**4.9**	**7.1**	4.1	-2.7	-0.7	1.8	-6
L_5	2.5	0.6	4.3	**4.4**	*4.4*	-0.6	1
$\chi^2(5)$[joint]	**20.20**	7.07	**15.60**	8.1	5.13	**15.5**	**14.9**

Vector autoregression is used to examine the impact of sentiment measures. The assumption of this model is that the expectation of regression residuals is independently and identically distributed (i.i.d). The endogenous variables included five lags of market returns, relevant market trading volumes, and sentiment measures. The exogenous variables include dummy variables of day-of-the-week, month-of-the-year, local holidays, 1987 financial crash effects, and five lags of the volatility proxy (we used GARCH(1,1) conditional variance of returns and VIX index as the volatility measures in the research). The trading volumes have been detrended from log volume. Robust standard errors (Newey and West, 1987)[67] are used to reduce the Heteroskedasticity of residuals.

The first model to be computed uses DJIA returns as the dependent variable for the period from 1984 to 1999 (results are presented in Table 4.13). The negative impact of 1^{st} day lag negative sentiment on DJIA returns is 4.8 basis points (1 basis point is a 0.01% change in return) and statistically significant at the 0.01 level; this is fully recovered within the 2^{nd} to 5^{th} days lag (with fourth-day lag statistically significant at 0.01 level). We then split this period into two: 1984-1991 and 1992-1999 to test the difference between normal and bull periods. The normal

period (1984-1999) includes the famous Black Monday crash but the bull period doesn't have any business downturns (according to NBER business cycles). The results showed that the impact of negative sentiment is more significant during the 1990s (negative 7.4 basis points 1^{st} day lag significance at 0.01 level) than it is during the 1980s (positive 7.1 basis points 4^{th} day lag significance at 0.05 level). The data period is further extended to the volatile 2000s (there are two recessions during the period 2000 to 2007). The results are similar to those in the 1980s; the 4.4 basis points lag five of negative sentiment is significant at 0.05 level. Following DJIA we looked at a cross-section of assets, aggregated stock market indices and crude oil futures and investigated the movement of the aggregates in three different markets (OMXC, SHCOMP and WTI). We extracted negative sentiment in news reports from local newswires and newspapers. The 1^{st} day lag negative sentiment in both OMXC and SHCOMP studies show significant coefficients (-9.5 and -6.0 respectively and significant at 0.01 level). In the WTI case, the effect is delayed to the 2^{nd} day lag, however, is still statistically significant (-8.5 at 0.01 level). The result confirms that there exists a linear relationship between negative sentiment and market returns in different markets.

4.7.2 Non-Linear Confirmation - Robust Locally Weighted Regression

Most researchers analyse the effect of investor sentiment extracted from media using parametric estimates. In order to visualise the non-linear relationship between the media factor and stock returns, we examined a semi-parametric approach. This approach comprises two steps: first, we conducted a parametric method that estimates the unexplained stock return residuals from Equation 4.1 excluding the lags of the sentiment measures; second, we obtained the semi-parametric estimates using a locally weighted regression method. Following the confirmation of linearity, we used one of the locally weighted regressions - Locally Weighted Scatterplot Smoothing (LOWESS) method to determine the non-linear relationships between sentiment and market returns.

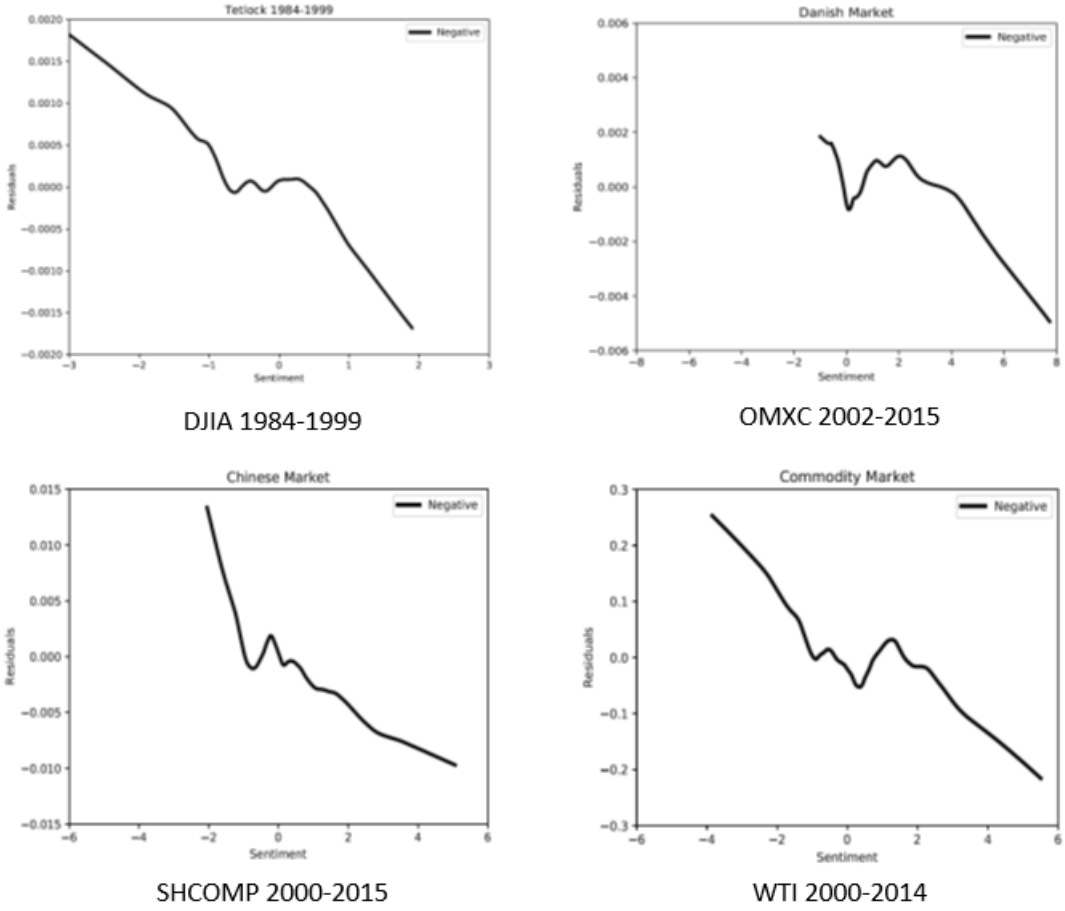


Figure 4.9: **Locally weighted regressions across markets: the x-axis denotes the standardised negative sentiment, and the y-axis denotes the market returns residuals.**

First, the lags of the negative sentiment measures ($L_5 S_t$) are excluded from the linear regression in Equation 4.1 and the residual is obtained from market returns. Second, LOWESS procedures are formed for the measures of negative sentiment. The x-axis denotes the standardised negative sentiment, and the y-axis denotes the market returns residuals.

The relationship graph that LOWESS plots, the smoothing parameter f[12] has been set to 0.4 and number of iterations[13] has been set to 1, similar previous studies (Tetlock, 2007; Yu, 2011;

[12]The smoother span. This gives the proportion of points in the plot which influence the smooth at each value. Larger values give more smoothness.

[13]The number of robustifying iterations which should be performed.

Garcia, 2013)[92][99][40]. Results are shown for four locally weighted regressions for measures of negative sentiment in four different markets (see Figure 4.9). All the measures (OMXC data only shows the relationship when the value of the negative sentiment is positive) demonstrated an adverse impact on market return residuals - the residuals monotonically decrease as negative sentiment increases with the exception of a short interval near the vertical axis. The graphs in Figure 4.9 show visualised non-linear relationships between negative sentiments and VAR model residuals. Apart from a small exception near the vertical axis, the DJIA returns decrease as negative sentiment increases over different periods. This remains true for the other three markets we explored (with the exception that the relation in the left hand-side of the graph for OMXC market doesn't hold when sentiment is lower than -1 standard deviation). These effects are more significant when the negative sentiment values are near the far positive or negative side. A robustness test was carried out by eliminating the outliers through winsorising the data. The results remained consistent throughout demonstrating that the estimations weren't being influenced by the presence of outliers.

It is found that both parametric and semi-parametric models indicate that sentiment proxy as a qualitative measure of investor sentiment has an impact on market returns (quantitatively); the computational results show the opposition of relationship between proxies and returns - high negative sentiment occurring with low returns and vice versa.

4.8 Case Study V – Firm-level Relationships between Sentiment and Return

In the previous three sections, we analysed the influence of investor sentiment on the market price through linear models, non-linear models and rolling models. In the next section, an introduction to the firm-level share price returns into the same model instead of the market price returns, revealing the impact of investor sentiment on the firm-level price is given.

4.8.1 Firm-level Linear Relationship Compared with Benchmark DJIA Index

First, the VAR model (Equation 4.1) is used to test the effect of negative sentiment on the return for each individual stock at the firm level. This test is divided into three parts. In the first part, the results of all the data from 2000 to 2015 are examined. In the second and third parts, the data period is cut off into two half - before and after 2017. The first half studies the results of the market expansion period and the second half studies the results of the market recession period.

Tables 4.14, 4.15 and 4.15 show the results of the three tests to study above. Through the summary of the tables, the analysis is followed:

During the full data period, we observed that 12 companies' share prices were affected by negative sentiment (Table 4.14). The negative sentiment has a statistically significant effect on both Apple and Exxon Mobil, based on the overall model; the effect on Apple is significant at 0.01 and Exxon Mobil at 0.1. From an industry perspective, six were affected by negative sentiment in the IT sector, three in the crude oil sector, two in the automotive sector and only one in the retail sector. From another perspective, it is needed to study the lagged impact of negative sentiment on stock returns. Negative sentiment had a significant negative impact on the one-day lag share price returns of five of the companies, having a significant negative impact on two companies' two-day lag share price returns, having a significant negative impact on three companies' three-day lag stock prices, and one company's four-day lag returns. Conversely, negative sentiment has had a significant positive effect on returns such as two-day lag of IBM and three-day lag of Toyota, as well as Intel's four-day lag. Taken together, the average effect of negative sentiment on 23 stocks, while not as statistically significant as they did on the benchmark Dow Jones index, is larger for chi-square value. In both cases, negative sentiment has a negative effect on the one-day lag. The difference is that reversion at the firm-level (four to five days lag) is slower than the index level (two to three days lag).

Table 4.14: **Negative Sentiment VAR Analysis: firm level news sentiment analysis in comparison with DJIA aggregated news, 2000 - 2015**

2000-2015																						
Negative Impact - Linear (VAR)																						
	L1			L2			L3			L4			L5			Chi-square			Industry		Market Cap
Apple	-12.8	*		-16.1	**		-0.9			5.4			15.1	*		20.66	***		IT		809.51B
Exxon	2.1			-8.5	**		4.3			-0.9			0.4			9.98	*		OI		326.00B
IBM	0.0			6.4	*		-3.2			-8.3	**		-0.5			8.79			IT		142.46B
HomeDepot	-6.5	**		3.5			-5.7			2.9			4.7			7.76			RE		233.38B
Microsoft	-10.1	*		1.0			6.4			8.3			3.7			7.76			IT		689.98B
Verizon	-6.7	**		-4.1			4.5			0.3			3.9			7.54			IT		208.07B
Ford	5.0			5.1			-11.5	**		-1.3			5.2			7.00			AU		42.41B
Conoco	1.9			-2.8			-5.7	*		-1.2			4.9			6.29			OI		65.48B
Toyota	-1.2			2.1			5.4	*		1.9			3.3			5.42			AU		203.22B
AT&T	0.7			2.1			-8.9	**		0.4			2.8			5.23			IT		266.74B
Intel	-4.8			-3.5			0.5			9.7	*		-4.1			3.94			IT		221.54B
Shell	-6.0	*		1.1			-0.3			0.2			1.8			3.60			OI		277.84B
Wal-Mart	0.4			3.9			-2.3			4.1			-4.4			6.26			RE		304.68B
Johnson&Johnson	-3.9			2.7			-3.1			2.2			-1.6			5.97			MA		353.09B
HP	-0.3			7.3			2.9			-2.8			7.1			4.07			IT		34.90B
Chevron	2.3			-2.2			-0.1			2.4			4.4			3.79			OI		218.98B
Pfizer	-5.1			-1.1			-2.7			-0.2			1.0			3.69			PH		208.51B
GE	-4.2			0.6			-5.6			1.5			-2.5			3.36			EE		132.99B
Boeing	-5.0			-2.7			-1.4			3.4			-3.0			3.34			AE		205.71B
Total	0.6			1.4			5.0			-3.7			1.1			3.28			OI		136.56B
Merck	-1.5			-3.5			-1.3			-0.6			-1.7			2.02			PH		152.24B
Cisco	-6.4			3.0			3.6			0.9			-1.2			2.01			IT		199.43B
BP	-2.8			0.4			0.0			0.4			-3.0			1.73			OI		121.62B
Average	-2.8			-0.2			-0.9			1.1			1.6			5.80					
Aggregated	-4.8			1.2			5.8			-2.5			-0.1			3.59					

Table 4.15: **Negative Sentiment VAR Analysis: firm level news sentiment analysis in comparison with DJIA aggregated news, 2000 - 2007**

	2000-2007							
	Negative Impact - Linear (VAR)							
	L1	L2	L3	L4	L5	Chi-square	Industry	Market Cap
Verizon	-12.8 **	-0.9	3.9	1.8	8.7 *	8.92	IT	208.07B
Ford	14.9 **	4.2	-12.4 **	-2.7	-3.8	8.39	AU	42.41B
Conoco	8.6 *	2.1	-8.4 **	-8.1 *	5.2	12.18 **	OI	65.48B
Apple	-13.2	22.0 ***	-0.6	2.3	16.9	15.64 ***	IT	809.51B
Total	2.0	0.9	7.7 **	-2.6	0.8	3.81	OI	136.56B
Johnson&Johnson	-8.9	2.8	-3.9	7.2 **	-7.4	11.38 **	MA	353.09B
Microsoft	-14.4	2.2	9.3	17.8 **	-0.5	8.85	IT	689.98B
IBM	6.0	-0.4	-0.6	-16.7 **	3.7	7.20	IT	142.46B
Toyota	0.2	5.5	4.3	-3.0	8.6 **	7.91	AU	203.22B
Chevron	0.8	-3.1	-3.1	-1.5	8.1 **	5.53	OI	218.98B
AT&T	-1.8	-8.4	-10.7	1.9	8.4	6.32	IT	266.74B
Exxon	1.6	-5.1	4.5	-6.4	3.5	6.32	OI	326.00B
Pfizer	-2.9	-9.1	-4.2	-4.7	2.3	5.23	PH	208.51B
Boeing	-0.8	-0.9	0.5	10.0	-9.0	4.71	AE	205.71B
HomeDepot	-6.8	7.3	-7.6	-0.6	4.5	3.85	RE	233.38B
Intel	1.4	-3.7	-14.2	14.8	-6.4	3.67	IT	221.54B
Wal-Mart	0.3	8.0	0.0	-0.5	-0.2	2.81	RE	304.68B
HP	0.8	5.4	1.9	-1.7	14.2	2.77	IT	34.90B
Cisco	-9.0	2.1	6.7	2.4	-1.1	1.42	IT	199.43B
Merck	-4.2	-1.1	2.5	-1.6	-2.2	1.37	PH	152.24B
Shell	-4.0	2.5	0.6	-0.6	0.5	1.32	OI	277.84B
GE	5.6	-2.7	-3.6	0.0	1.3	1.28	EE	132.99B
BP	0.7	-3.8	-1.1	1.6	-0.2	1.13	OI	121.62B
Average	-1.6	1.1	-1.2	0.4	2.4	5.74		
Aggregated	1.3	-0.7	3.9	-0.8	-1.0	0.98		

Table 4.16: **Negative Sentiment VAR Analysis: firm level news sentiment analysis in comparison with DJIA aggregated news, 2008 - 2015**

2008-2015																										
Negative Impact - Linear (VAR)																										
	L1			L2			L3			L4			L5			Chi-square			Industry		Market Cap				
HomeDepot	-8.0	**		-0.5			-5.7			4.0			3.4			6.55			RE		233.38B				
Pfizer	-8.2	**		2.6			-2.4			0.8			-1.3			5.15			PH		208.51B				
GE	-13.8	*		3.7			-7.2			2.0			-4.3			6.67			EE		132.99B				
IBM	-2.9			11.8	***		-5.4			-3.2			-2.5			10.82	*		IT		142.46B				
AT&T	4.6			12.2	**		-4.0			2.4			-0.4			8.63			IT		266.74B				
Exxon	2.5			-13.2	**		4.2			7.1			-2.6			10.87	*		OI		326.00B				
Verizon	0.6			-7.2	*		5.7			-1.3			-1.7			3.87			IT		208.07B				
Wal-Mart	0.3			-0.8			-5.3	*		7.6	***		-8.1	***		14.15	**		RE		304.68B				
Apple	-12.4			-3.2			2.4			16.8	*		8.6			5.75			IT		809.51B				
Chevron	3.8			-1.4			4.2			7.2	*		0.9			4.08			OI		218.98B				
Intel	-7.0			-4.1			9.6			9.7	*		-2.6			6.67			IT		221.54B				
Toyota	-2.6			-2.3			6.6			8.1	*		-2.7			5.24			AU		203.22B				
Conoco	-3.8			-6.9			-3.4			6.1			5.6			5.94			OI		65.48B				
Boeing	-10.5			-4.4			-3.7			-4.8			2.2			5.39			AE		205.71B				
Ford	-5.8			6.2			-12.7			-4.0			10.5			4.69			AU		42.41B				
BP	-6.6			5.4			1.5			0.7			-5.7			3.26			OI		121.62B				
HP	0.6			9.0			4.2			-2.4			3.7			2.78			IT		34.90B				
Merck	0.1			-5.1			-5.1			-0.1			-1.6			2.68			PH		152.24B				
Shell	-7.4			0.4			-1.9			1.3			3.5			2.53			OI		277.84B				
Johnson&Johnson	-0.6			2.9			-2.0			-2.1			2.4			2.50			MA		353.09B				
Microsoft	-7.3			-1.6			3.7			-0.5			4.2			1.52			IT		689.98B				
Total	-1.5			2.5			2.3			-5.1			1.5			1.23			OI		136.56B				
Cisco	-3.8			3.9			1.8			-0.1			-0.3			0.94			IT		199.43B				
Average	-3.9			0.4			-0.6			2.2			0.6			5.30									
Aggregated	-13.2	*		0.2			4.9			-2.6			-0.7			5.03									

In the same way, we used this method to study the impact of negative sentiment on the stock market during its expansion (Table 4.15). Relative to the full data range, only ten stocks in the expansion period were affected by negative sentiment. Among them, if the statistics of overall model is calculated, Apple is significant at 0.01 level, while Conoco and Johnson & Johnson are significant at 0.05 level[14]. Most of the industries that are affected by the impact of negative sentiment were still concentrated in the IT, crude oil and automotive industries. Slightly different is that no retail sector has been affected, but Johnson & Johnson entered the list as a representative of the manufacturing industry in the expansion period. From the view of the lagged impact of negative sentiment on the stock prices, one with one-day lag, one with two-day lag, one with three-day lag and two with four-day lag that were negatively affected by negative sentiment. On the contrary, one on one-day lag, one on two-day lag, one on three-day lag, two on four-day lag and three on five-day lag were positively affected by negative sentiment. In general, the negative sentiment still has no significant statistical impact on both the firm-level and the index-level. However, comparing chi-square values, the firm-level is relatively stable, while the index-level dropped from 3.59 to 0.98.

To study the recession, the 2008-2015 data is inserted into the model (Table 4.16). There are 12 stocks in this range affected by negative sentiment. The results of the industry research have changed, the most affected is the IT industry - there are five IT stocks in the list. Crude oil industry reduced to two stocks, the automotive industry to one. Retail increases to two stocks, pharmaceutical to one, electrical engineering to one. Taking a closer look at the lagged impact, the negative impact of negative sentiment is reflected in the one-day lag of three stocks, two stocks lagging two days, one stock lagging three days and one stock lagging five days. The positive impact of negative sentiment is reflected in two stocks with two-day lag and five stocks with four-day lag. On average, only negative sentiment has a significant negative impact on the

[14]It is known that the risk of type-I error increases with multiple tests. It is believed that the significance level is the probability value as an indication of the weight of evidence against the null hypothesis rather than making a decision to reject the null hypothesis. The risk of such error has also been lowered by reducing the sampling and modelling uncertainties that have been discussed earlier in Chapter 2.

first-day lag of the index level returns. Compared with the former two cases, the chi-square value at the firm-level is still stable. The value at the index-level has risen over the previous two.

4.8.2 Empirical Testing over Rolling-Window Periods

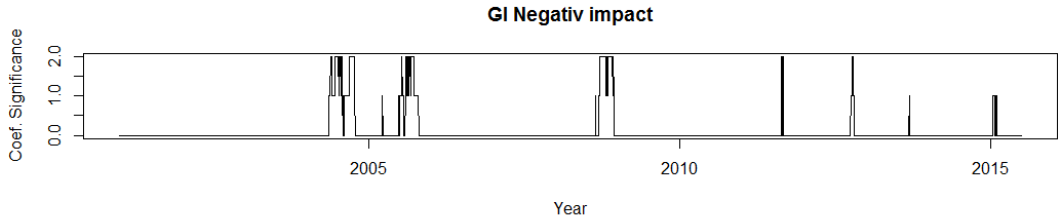
The model used previously was to test the sentiment analysis hypothesis of the market and the firm level with the VAR model. In the previous tests, it is found that sentiment has varying degrees of impact on returns at both the market and the firm level. The following section will be focused on time-varying data for the regression test.

The time-varying test will be applied to the one-year rolling window. The goal is to test whether sentiment has an impact on returns when time varies. With the rolling model, one can detect the impact of sentiment on indices (or company stocks) for any period of one-year length. Based on the data obtained, a daily rolling model of 15 years data is conducted using 250-day windows starting on the first day of 2000. The results are shown in graphs, each contains a line indicating the level of the significance of the negative sentiment impact. If the result on a day is 1 on the figure, it means that the negative sentiment impact on company's stock price return on the day is statistically significant at 0.05 level; if the result is 2, the significant level is 0.01.

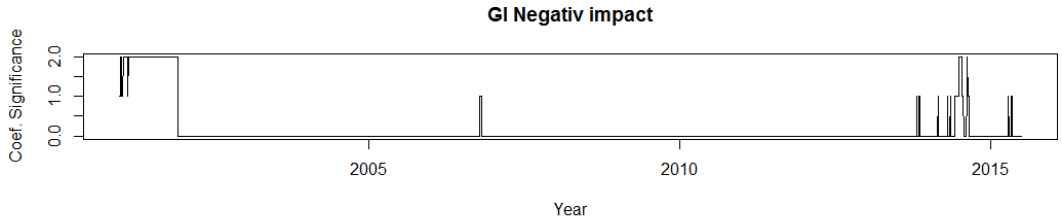
From Figure 4.10[15], we observed that there are only two significant impacts over 200 days a year, Apple in 2001 and Ford in 2004. In general, the number of days that had a significant impact by 2007 was greater than after 2007. On average, the number of days with significant impact before 2007 was 30 days/year, compared with 19 days/year after 2007. Considering the sample data covers many companies in different industries, at a firm level sentiment analysis, the impact of sentiment is present and the effect is selective.

[15]Figures follow the order of chi-square scores in Table 4.14.

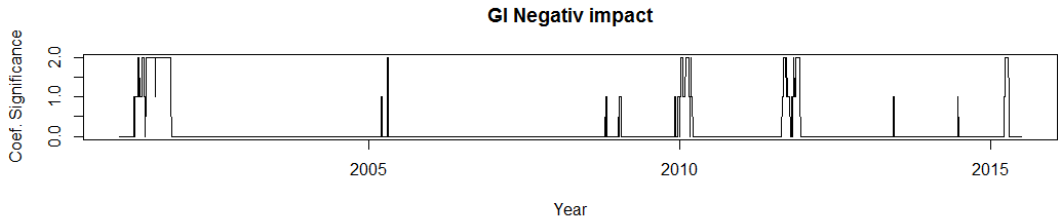
DJIA



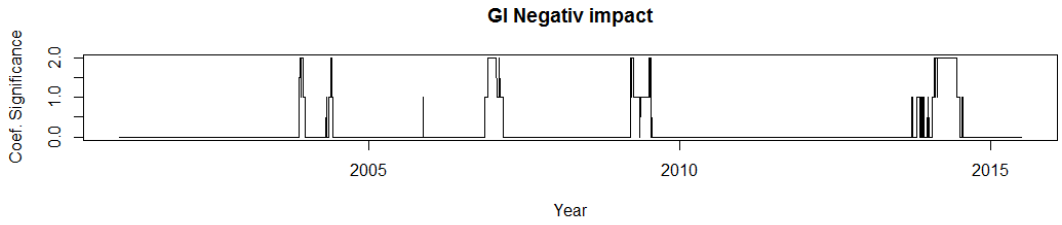
(a)
Apple



(b)
Exxon



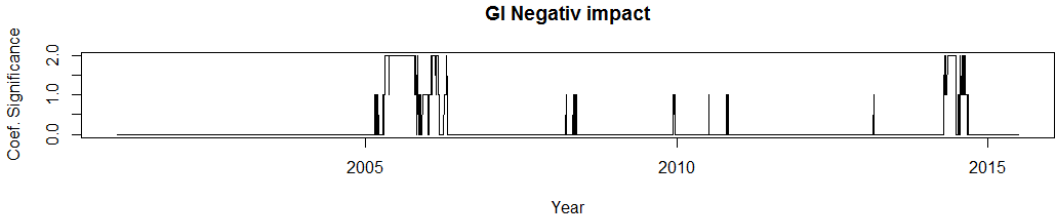
(c)
IBM



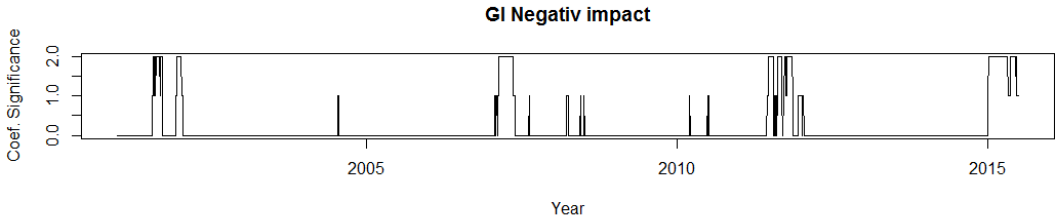
(d)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year *(cont.)***

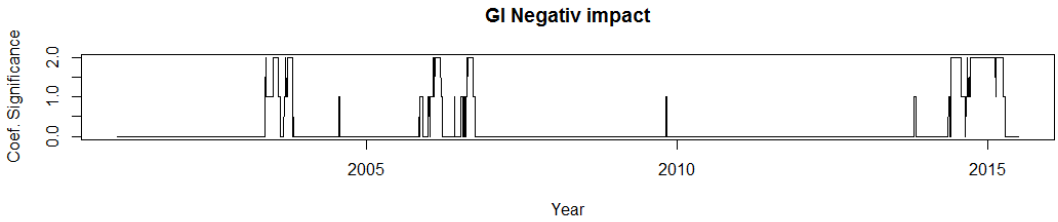
HomeDepot



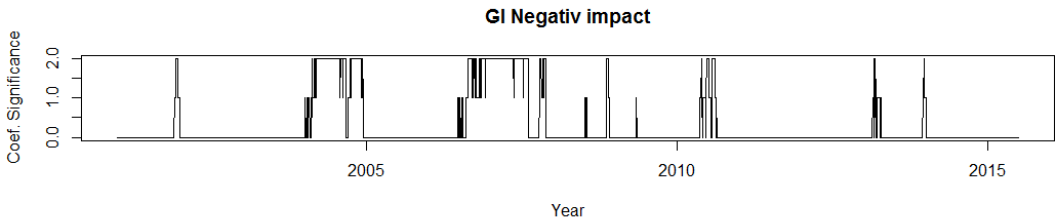
(e)
Microsoft



(f)
Verizon



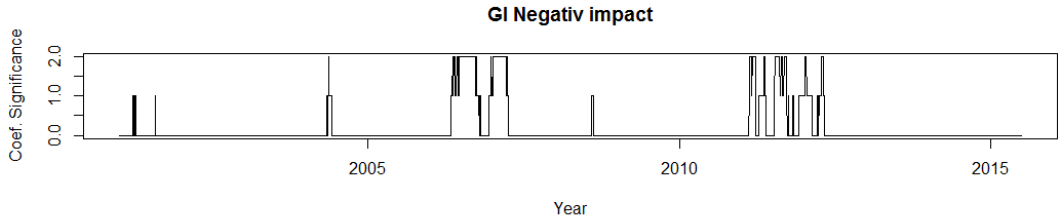
(g)
Ford



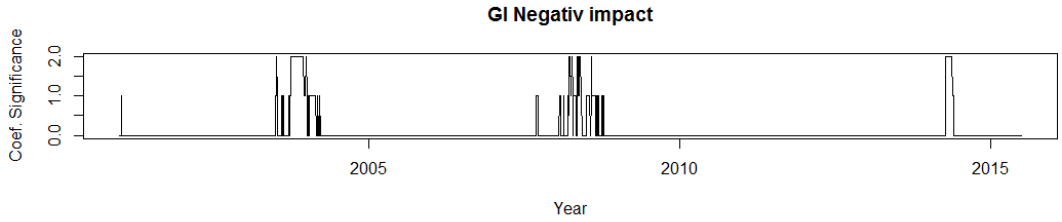
(h)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year *(cont.)***

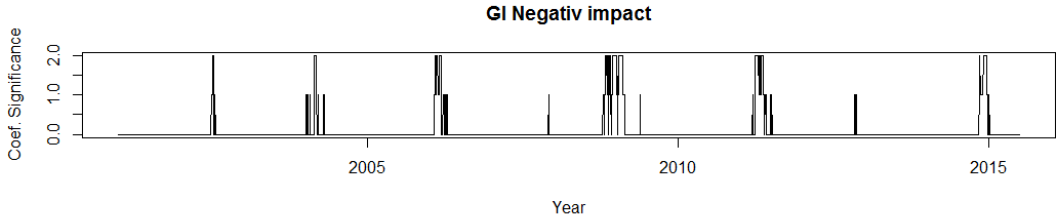
Conoco



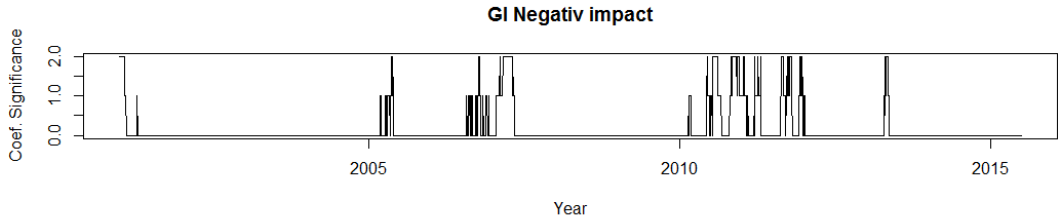
(i)
Toyota



(j)
AT&T



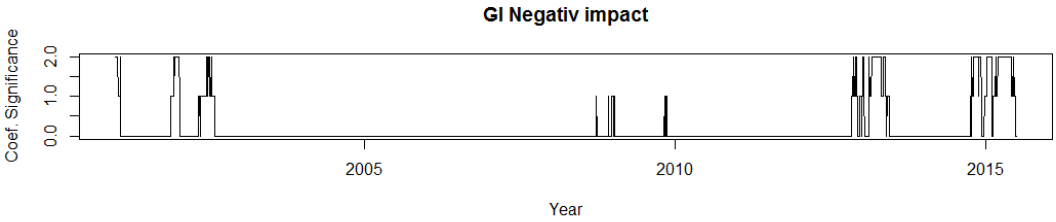
(k)
Intel



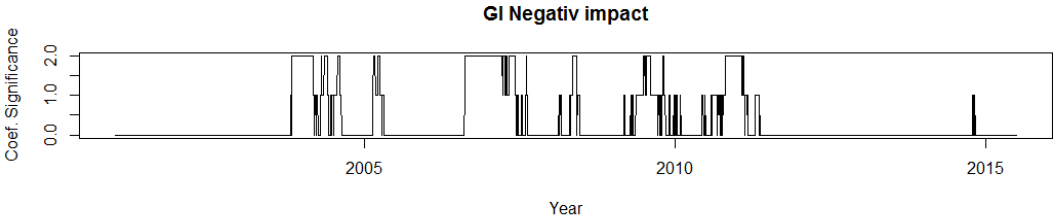
(l)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year *(cont.)***

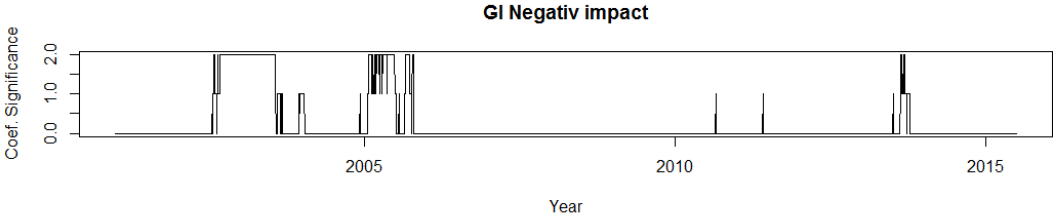
Shell



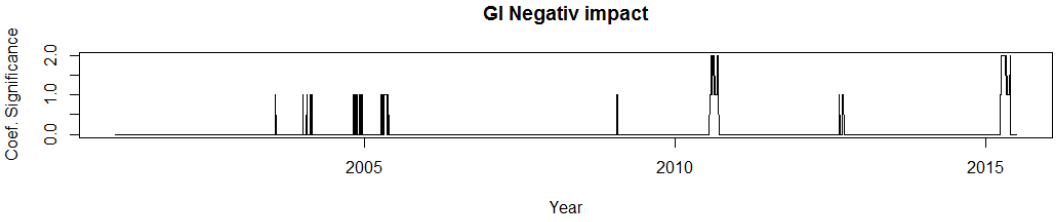
(m)
Wal-Mart



(n)
Johnson&Johnson



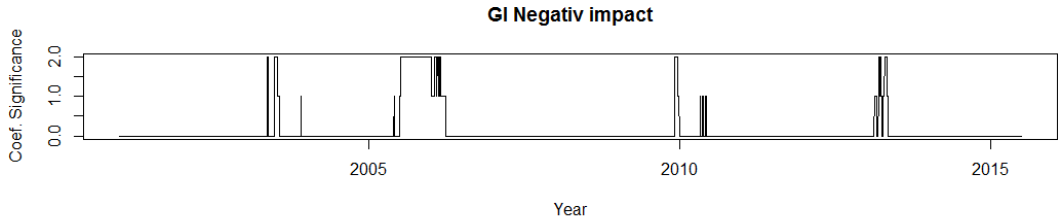
(o)
HP



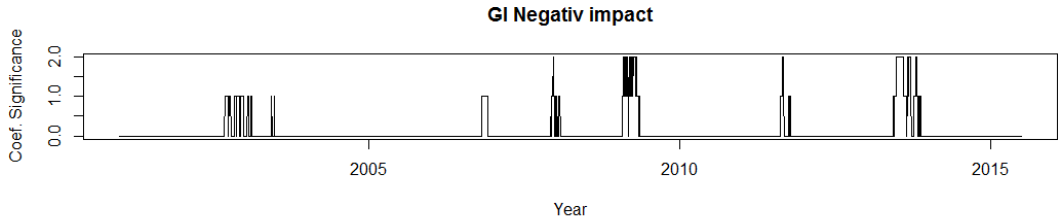
(p)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year *(cont.)***

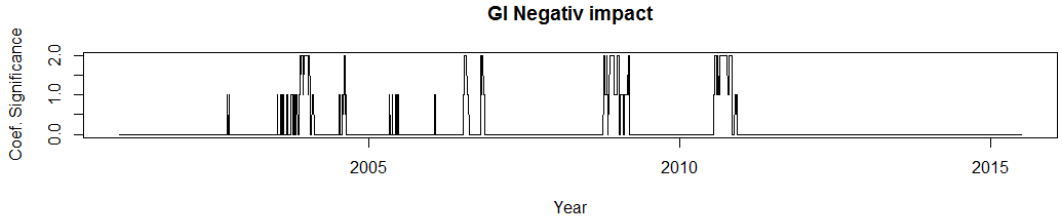
Chevron



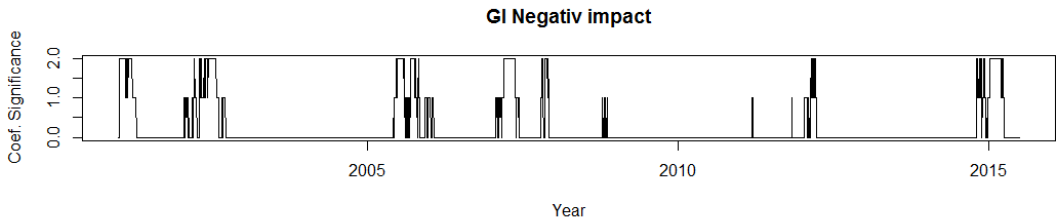
(q)
Pfizer



(r)
GE



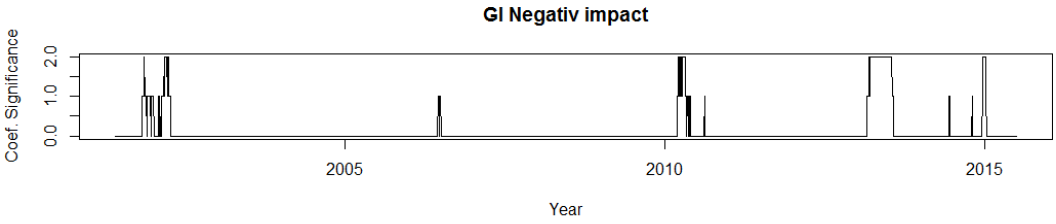
(s)
Boeing



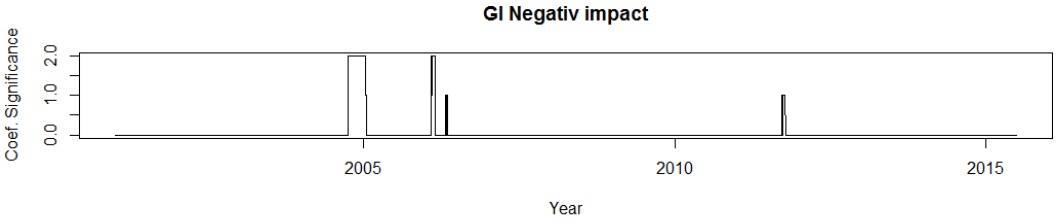
(t)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year *(cont.)***

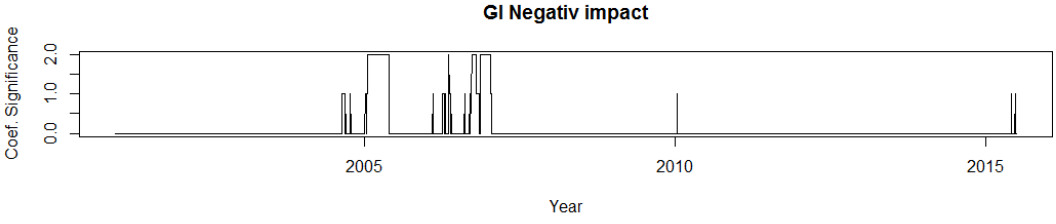
Total



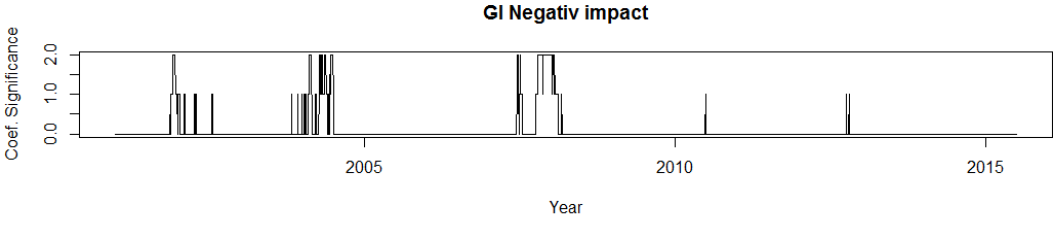
(u)
Merck



(v)
Cisco



(w)
BP



(x)

Figure 4.10: **Number of Days that Negative Sentiment Impact is Significant in a Year in Rolling Windows: firm level news sentiment analysis in comparison with DJIA aggregated news. Note: in 250-day window rolling regressions, results start from the second year**

4.8.3 Empirical Testing - Correlation between Positive and Negative Sentiments over Rolling Windows

Prior to this section, only the negative glossary of the GI dictionaries is used to study investor sentiment. However, it is crucial to analyse the correlation between positive and negative sentiments for this research and to investigate the influence of investor sentiment during expansion or recession using both sentiments.

The data in Table 4.17 and 4.18 shows the correlation coefficients between positive sentiment and negative sentiment in each company's sentiment analysis data for each year and whether they are statistically significant. The expansion period and decline period are also divided to compare the results before and after a crisis. Before 2007, one-half (97/184) of the coefficients were positive and statistically significant, and only two of them were negatively correlated and significant. After 2007, 16/184 coefficients were positively correlated and significant, while negatively correlated and significant figures increased to 8. According to the annual average, the average annual correlation coefficient before 2007 is 20%, and after 2007, the correlation coefficient is 1%.

4.8.4 Empirical Testing - Comparison between Benchmark Index and Firm Level using LOWESS Results

In the previous section, we conducted a non-linear analysis of the negative sentiment impact on return residuals at the index-level. The result shows that the negative sentiment affects the return residuals in a phased manner. When the negative sentiment is low (< -1), the corresponding residual value is larger; when the negative sentiment is higher (> 1), the corresponding residual is petite; while the negative sentiment is close to 0, the residual is also floating near the zero.

In this section, we use the same semi-parametric model to test the relationship between firm-level return residuals and negative sentiment. During the analysis, a graph for each company's data is generated. According to the expected LOWESS smooth curve (see Figure 4.11), if semi-parametric images appear in the upper left and lower right quadrants, and there is irregular

Table 4.17: **Correlation analysis between *positive* and *negative* sentiment: firm level news sentiment analysis in comparison with DJIA aggregated news 2000-2007**: bold number denotes statistically significant correlation at 0.01 level.

	2000	2001	2002	2003	2004	2005	2006	2007
Apple	7%	-4%	-4%	9%	3%	-3%	-6%	-16%
AT&T	**65%**	**69%**	**68%**	**55%**	**71%**	**48%**	**39%**	**21%**
Boeing	**21%**	11%	**19%**	7%	**34%**	**27%**	**37%**	15%
BP	**29%**	**23%**	10%	-5%	8%	14%	3%	-2%
Chevron	7%	-2%	**31%**	**18%**	13%	0%	11%	**-24%**
Cisco	**34%**	**33%**	**17%**	**27%**	**33%**	16%	**19%**	-3%
Conoco	**20%**	**20%**	1%	5%	4%	-2%	10%	-6%
Exxon	4%	11%	13%	10%	-3%	7%	-1%	4%
Ford	15%	12%	**16%**	4%	12%	**22%**	**24%**	11%
GE	**33%**	**34%**	**37%**	**35%**	**44%**	**51%**	**47%**	13%
HomeDepot	**22%**	9%	-3%	4%	12%	**24%**	13%	**-18%**
HP	**20%**	**20%**	9%	**24%**	**27%**	**26%**	**20%**	-8%
IBM	**31%**	-3%	**40%**	13%	15%	**32%**	13%	1%
Intel	**17%**	**34%**	**21%**	**25%**	**20%**	**30%**	8%	**17%**
Johnson&Johnson	**18%**	**36%**	**22%**	14%	**20%**	**30%**	**22%**	-1%
Merck	14%	**27%**	15%	11%	16%	**27%**	**34%**	**25%**
Microsoft	7%	9%	10%	14%	**38%**	**32%**	**45%**	**31%**
Pfizer	**42%**	**37%**	**31%**	**45%**	**29%**	**39%**	**41%**	-3%
Shell	9%	16%	11%	-11%	-5%	14%	**23%**	14%
Total	4%	-1%	**23%**	11%	9%	14%	**22%**	3%
Toyota	**26%**	16%	14%	**31%**	9%	**18%**	**32%**	**24%**
Verizon	**53%**	**53%**	**63%**	**58%**	**66%**	**52%**	**40%**	**36%**
Wal-Mart	**26%**	**19%**	**18%**	**24%**	**26%**	**26%**	**22%**	5%
Summary	23%	21%	21%	19%	22%	24%	23%	6%
	20%							
Aggregated	19%	32%	23%	57%	44%	42%	61%	-4%

Table 4.18: **Correlation analysis between *positive* and *negative* sentiment: firm level news sentiment analysis in comparison with DJIA aggregated news 2008-2015**: bold number denotes statistically significant correlation at 0.01 level.

	2008	2009	2010	2011	2012	2013	2014	2015
Apple	17%	-5%	-7%	2%	-1%	-7%	4%	-3%
AT&T	5%	14%	15%	9%	13%	2%	-3%	-9%
Boeing	**20%**	**31%**	**-16%**	-12%	-7%	**-23%**	-15%	-6%
BP	-16%	3%	-3%	5%	4%	8%	-3%	-1%
Chevron	-9%	9%	1%	5%	7%	13%	10%	5%
Cisco	7%	11%	-2%	-5%	6%	**17%**	**27%**	13%
Conoco	-5%	11%	6%	4%	-10%	8%	6%	3%
Exxon	**-20%**	8%	15%	14%	13%	-2%	**20%**	-1%
Ford	2%	**22%**	1%	-7%	-2%	-6%	9%	-10%
GE	6%	**18%**	1%	11%	**16%**	**20%**	**46%**	20%
HomeDepot	**-22%**	-13%	2%	-15%	**-20%**	-12%	1%	-14%
HP	**-19%**	-1%	1%	-7%	-13%	2%	14%	-12%
IBM	-12%	-13%	-9%	-11%	5%	12%	6%	-21%
Intel	15%	0%	-10%	14%	-3%	10%	-15%	-5%
Johnson&Johnson	0%	-4%	1%	**-20%**	6%	2%	2%	-10%
Merck	-13%	-4%	-13%	0%	0%	12%	-10%	-21%
Microsoft	**29%**	**24%**	-7%	-4%	-8%	**-19%**	-14%	**-31%**
Pfizer	-1%	-1%	**-18%**	-11%	-1%	1%	-12%	**-24%**
Shell	-6%	1%	-4%	-7%	12%	10%	12%	**27%**
Total	-1%	2%	-1%	-3%	9%	7%	8%	-7%
Toyota	-9%	4%	2%	-2%	-1%	15%	**19%**	-2%
Verizon	**51%**	**44%**	10%	-15%	6%	14%	15%	-6%
Wal-Mart	9%	2%	**-17%**	-10%	**19%**	-7%	-6%	-3%
Summary	1%	7%	-2%	-3%	2%	3%	5%	-5%
	1%							
Aggregated	20%	47%	-32%	2%	26%	-18%	-10%	2%

Table 4.19: **Negative impact - non-linear (LOWESS)**: when the graph of a company's quadrant section is completely/nearly similar to the graph of indices, it is believed that the impact of negative sentiment on residuals meets expectations, and a value of 1 is given; if the quadrant is incomplete but roughly similar, a value of 0.5 is given; if the quadrant is entirely different, the value is 0. "++" denotes the quadrant of positive residual vs positive sentiment; "+-" denotes the quadrant of positive residual vs negative sentiment; "-+" denotes the quadrant of negative residual vs positive sentiment; "–" denotes the quadrant of negative residual vs negative sentiment.

	2000-2015						2000-2007						2008-2015				
	+-	++	-+	–	Sum		+-	++	-+	–	Sum		+-	++	-+	–	Sum
Boeing	1	1	1	1	4	Boeing	1	1	1	1	4	Boeing	1	1	1	1	4
Johnson	1	1	1	1	4	Johnson	1	1	1	1	4	BP	1	1	1	1	4
Merck	1	1	1	1	4	Merck	1	1	1	1	4	Ford	1	1	1	1	4
BP	0.5	1	1	1	3.5	BP	1	1	0	1	3	Johnson	1	1	1	1	4
Exxon	1	1	0	1	3	HomeDepot	0	1	1	1	3	Apple	1	1	1	0	3
Total	1	1	0	1	3	Intel	1	0	1	1	3	Chevron	0.5	1	0.5	1	3
Wal-Mart	1	1	0	1	3	Pfizer	1	1	0	1	3	Merck	1	1	0	1	3
Cisco	1	0	0	1	2	Shell	1	1	0	1	3	Pfizer	1	1	0	1	3
GE	1	0	0	1	2	Total	1	1	0	1	3	Shell	1	1	0	1	3
HP	1	0	0	1	2	Cisco	1	0	0	1	2	Cisco	0.5	1	0	1	2.5
Intel	1	0	0	1	2	Ford	1	0	0	1	2	Microsoft	0.5	0	1	1	2.5
Pfizer	1	0	0	1	2	GE	1	0	0	1	2	Conoco	0.5	0	1	0.5	2
Shell	1	0	0	1	2	Wal-Mart	1	0	0	1	2	Exxon	0	1	1	0	2
Toyota	1	0	0	1	2	AT&T	0	0	0	1	1	GE	1	0	0	1	2
Chevron	0	1	0.5	0	1.5	Exxon	0	0	0	1	1	IBM	0	1	1	0	2
Ford	1	0	0	0	1	Toyota	0	0	0	1	1	Intel	0	1	1	0	2
HomeDepot	0	1	0	0	1	Apple	0	0	0	0	0	Toyota	0	1	1	0	2
IBM	0	0	0	1	1	Chevron	0	0	0	0	0	Wal-Mart	0	1	0	1	2
AT&T	0	0	0	0.5	0.5	Conoco	0	0	0	0	0	HomeDepot	0	0.5	0	1	1.5
Apple	0	0	0	0	0	HP	0	0	0	0	0	AT&T	0	0	0	1	1
Conoco	0	0	0	0	0	IBM	0	0	0	0	0	Total	0	0	1	0	1
Microsoft	0	0	0	0	0	Microsoft	0	0	0	0	0	Verizon	0	1	0	0	1
Verizon	0	0	0	0	0	Verizon	0	0	0	0	0	HP	0	0.5	0	0	0.5
DJIA	1	1	1	1	4	DJIA	1	1	1	1	4	DJIA	1	1	1	1	4

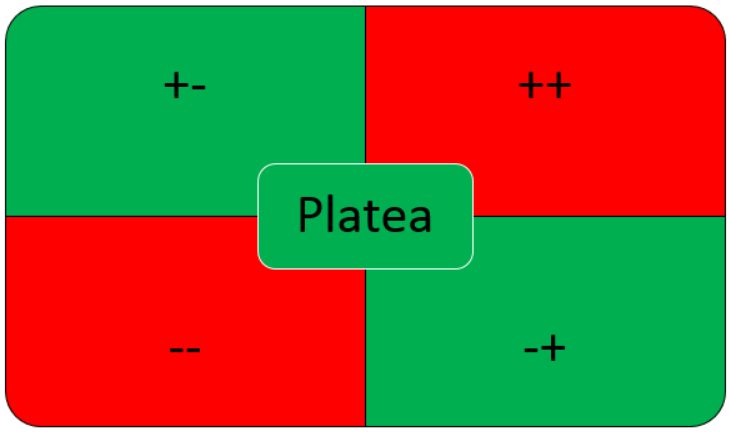


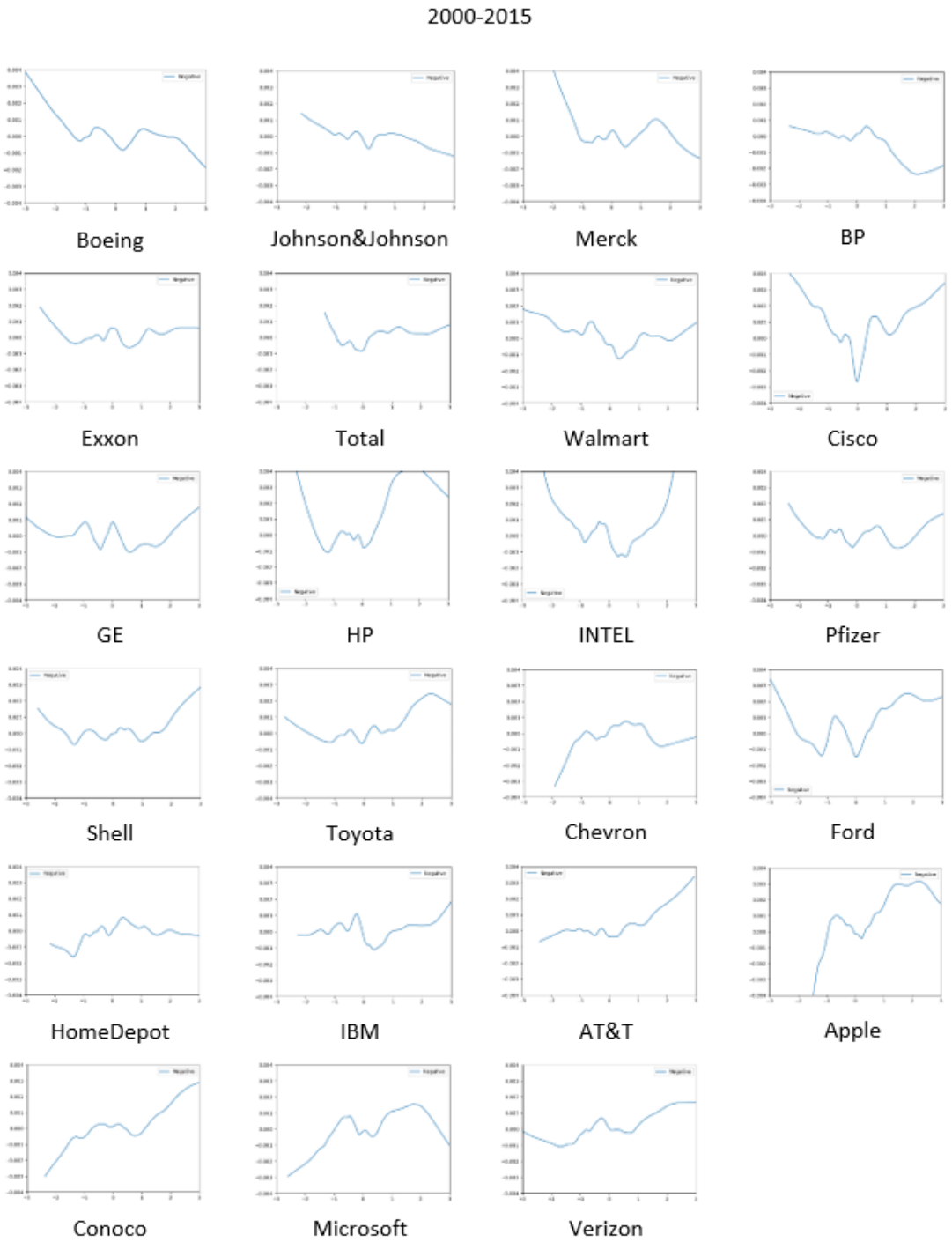
Figure 4.11: **The expectation of a LOWESS smoothed curve shown in four quadrants**: smoothed curve of negative sentiment vs residual in "+-" and "-+" and fluctuations in "Platea" are expected. "++" denotes the quadrant of positive residual vs positive sentiment; "+-" denotes the quadrant of positive residual vs negative sentiment; "-+" denotes the quadrant of negative residual vs positive sentiment; "–" denotes the quadrant of negative residual vs negative sentiment.

fluctuation in the middle "Platea" part, the image is considered to be expected, and the influence of negative emotions on residuals can be observed.

We summarise the four quadrants for each company(see Table 4.19), in three different periods, and give each company a total value for more natural sorting and comparing. It seems that most of the traditional industries (crude oil, pharmaceuticals, retail, etc.) have higher scores, while emerging industries (such as IT, communications, etc.) score lower. This result is generally applicable in three periods. Company's LOWESS graph is shown in Figure 4.12 (ranked according to the score in the previous table). Boeing, Johnson & Johnson, and BP's graphs are all similar to the Dow Jones Industrial Average in the three periods, while it is difficult to identify similarities from graphs of other companies. Some IT companies are in opposite states.

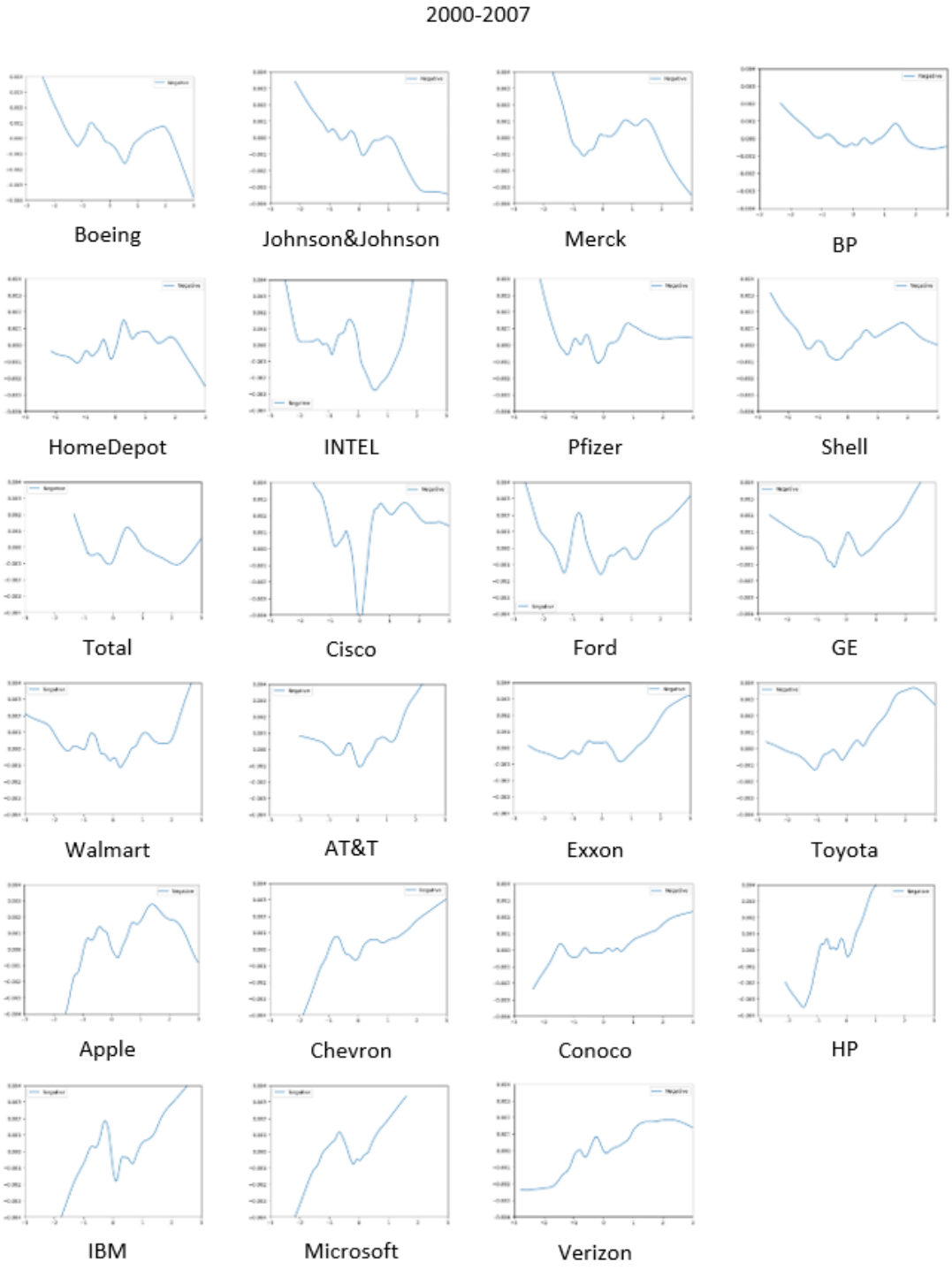
4.9 Conclusions

In these case studies, the impact of negative investor sentiment has been tested by extracting negative word frequency from legacy media sources using the BoW model. The results show



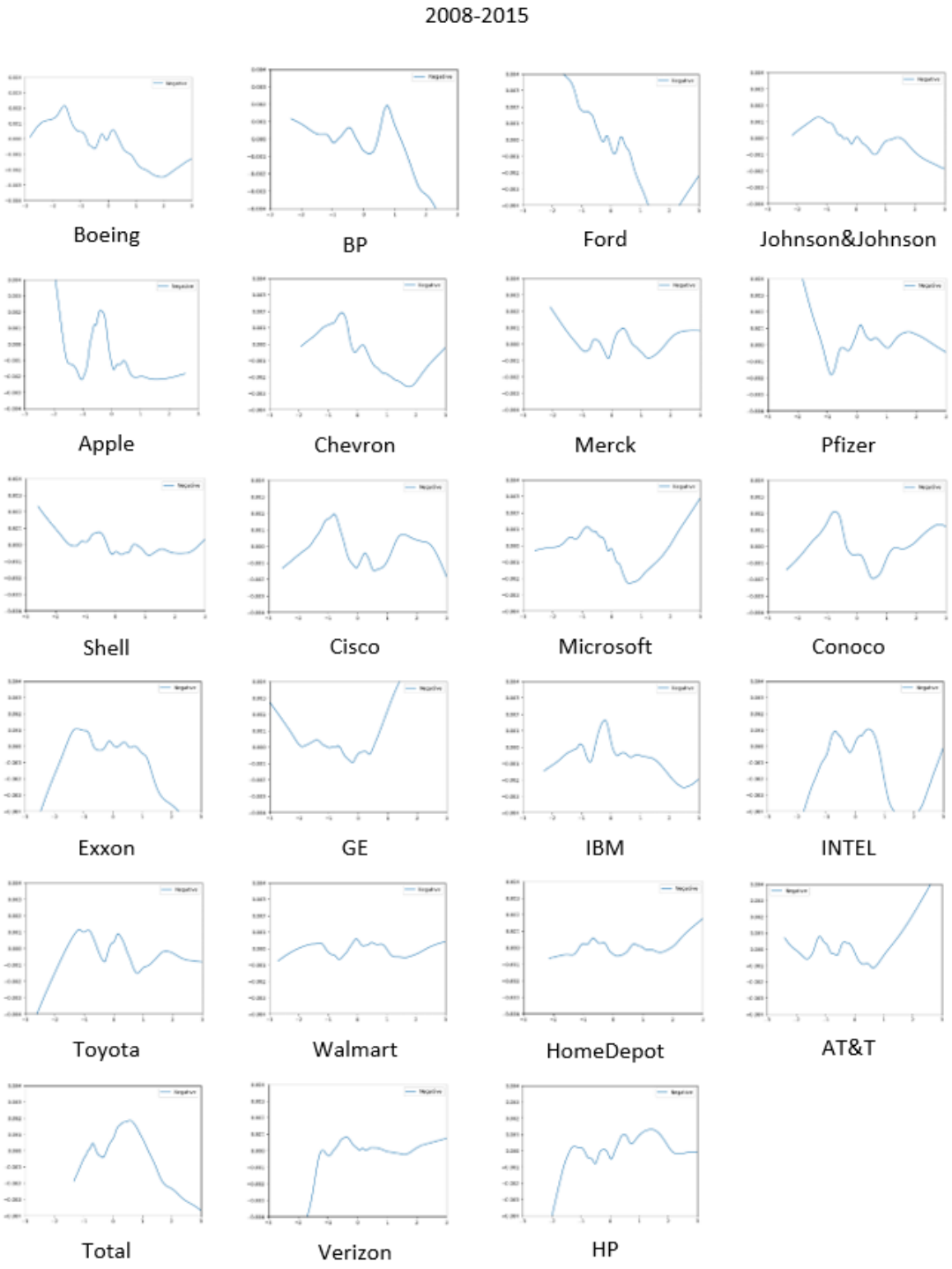
(a)

Figure 4.12: **LOWESS graphs of the 23 individual firms 2000-2015** ***(cont.)***



(b)

Figure 4.12: **LOWESS graphs of the 23 individual firms 2000-2007 (Before Crisis)** ***(cont.)***



(c)

Figure 4.12: **LOWESS graphs of the 23 individual firms 2000-2007 (After Crisis)**

statistical significance in three different markets. The benchmark study was entirely a replica work of the Tetlock (2007)[92] paper. Over a 15-year period (from 1984 to 1999) and its extended period (until 2007) the negative sentiment extracted from the opinion column on *Wall Street Journal* has a significant impact on Dow Jones Industrial Average index. The 1^{st} and 4^{th} day lag have respective impacts of 4.8 basis points negative and 4.9 positive in the 1984-1999 period. We found that the impact of negative sentiment is more significant in a bull period of the 1990s than in the non-bull period of the 1980s. The impact of negative sentiment in news reports from WSJ and NYT is different from commentaries from 2000 to 2015. Comparably, from 2000 to 2015, investor sentiment in the Danish and Chinese markets played an important role. The 1^{st} day lag of negative sentiment always has a statistically significant negative impact on the stock returns. The difference is that the impact is recovered on the 5^{th} day in Denmark but much quicker on the 2^{nd} day in China. Looking at the Granger causal relationship, the impact of negative sentiment on stock prices in the US and Chinese markets are significant at the 0.01 level however in the Danish market it's less significant. In the firm-level analysis, only about half of company returns affected by negative sentiment (this is true at three different times). Before and after the depression period, the company's aggregation was different (oil companies were generally not affected by sentiment during the growth period, but after the recession period, the sentiment impact increased correspondingly; it is also true for some IT companies).

A two-stage analysis has also been conducted - parametric and semi-parametric regressions to explore the relationship between media sentiment and market returns across DJIA, OXMC, SHCOMP and US firms. In particular, although observed that the estimates of the effect of negative sentiment on returns are broadly consistent with the parametric coefficients (linear regressions - VAR results indicate that negative sentiment generally has a pessimistic impact on the market the first two days after the news release and a reversal conclusion within five days), the results of semi-parametric procedures, using the non-linear method (locally weighted regression), show differences. The sentiment impact is not always true in firm-level returns. Sentiment forecasts remarkably large and tenacious slumps in the returns of small stocks, indicating sentiment estimates investors' opinions. Overall, the tests here recognise return patterns consistent

with the hypothesis that the sentiment estimates of words in the text corpus are valid sentiment indicators. After joining the firm-level tests, VAR models, rolling models, and the LOWESS model are applied to examine the effects of sentiment. Various tests have shown that the impact of investor sentiment on the firm-level stock market is not always observed.

Chapter 5

Conclusion

This thesis implemented a method to test the impact of the media and related sentiment on the stock market. Compared to other works (Tetlock, Garcia, etc.), the effect of sentiment on firm-level stock price volatility is examined in more detail. In recent years, content analysis and machine learning methods have become a favourite way of measuring financial news content. The impact of the news media and announcements on financial markets has been discussed widely. Literature research using news and text has been steadily growing. The role of textual sentiment in financial markets has been of interest to financial researchers for nearly a decade, especially since 2007. The literature on codes of conduct argues that human beings are emotionally and psychologically biased which may cause them to behave in an irrational manner. By analysing news and other sources of information, it may be possible to get a general idea of how investors think and feel. Researchers focus on the relationship between textual sentiment and company fundamentals or market variables. These sentiment reactions may cause investors to overestimate or underestimate the value of a particular asset beyond its "true" (underlying) value. This can be used to determine whether investors perceive assets as overvalued or undervalued.

It has been suggested that the tone of the news can be estimated by counting the number of sentiment words appearing in an article. Financial studies have been undertaken of text content and intonation from news articles and social media to earnings release and corporate disclosure.

Text sentiment in news reports may be a more objective reflection of the state of business or other objects. It may also convey new information and have fundamental explanations of asset prices. The tone of news estimates has been shown to predict financial returns. Newspapers have the advantage of being more frequent than companies disclosing and covering many events. News reports are often more objective and require less pre-processing time than Internet publishing. This thesis extracts sentiment from firm-specific news coverage and examines its impact on asset prices and market activity. The predictability of news in the financial markets is mainly explained in the financial literature in two ways. News reports do not fully incorporate price and market information, or new attitudes and beliefs that affect investors motivate responses. The application of text analysis methods in financial modelling has been diversified. The work of this thesis also contributes to this growing field of research.

This thesis analysed the application of content analysis in finance and the role of textual sentiment in different markets. This chapter is mainly to summarise the results.

By testing the benchmark DJIA index, a relatively smaller European index and an emerging index and by examining firm-specific sentiment using data from the United States, Denmark, and China through the parametric models (VAR model and rolling regression) and semi-parametric models (LOWESS models), this thesis examines the relationship between textual sentiment and market-level returns and correlation between the returns of the firm's shares.

This chapter provides concluding comments on each subject and suggests areas that future research may need to consider. The remainder of this chapter will discuss the contributions of this thesis and the background of these findings in the interdisciplinary area of financial content and sentiment analysis (Section 5.1), and describes the limitations of work and potential future work (Section 5.2).

5.1 Contributions

The main contribution of this thesis is to develop an approach to compute sentiment time series from a series of news articles or annotated texts. This sentiment variable can then be aggregated

with the financial time series in the statistical model to examine any correlation between news and financial markets. Chapter 3 first described the approach framework that allows any text corpus to be imported, and can also import time series. Specially designated methods allow different data sets to be used with the approach. The approach allows different types of news to be used to estimate the effects of sentiment variables and news in different financial markets. This allows less subjective analyses but also facilitates the investigation of the relationship between different text types and time-series data. Second, the approach estimates the impact of sentiment variables and news at the company level using integrated vector autoregression, rolling regression, and LOWESS regression analysis.

The case study presented in Chapter 4 assesses the implementation of systems and methods for building sentiment proxies from news texts. In order to test any correlation between financial markets and sentiment variables, many time series models are estimated in the modelling part of the approach to examine the impact of news on returns. The VAR specification is evaluated with control variables that account for market anomalies and seasonal variations. These control variables take into account past market returns, the impact of trading volume, volatility, and calendar effects. Control variables are attempts to isolate the effects of sentiment by including variables that also explain the same information. Sentiment effects are considered consistent because VAR models are estimated by the approach step by step, including controlling variables one at a time to see if there is any confusion with the sentiment variables. As far as the Dow Jones Industrial Average is concerned, the results show that the Dow Jones Industrial Average has a -4.7 basis points impact on one standard deviation change (increase) in the negative sentiment of the Dow Jones Market Index. The Danish local economic news (Danish version to English translation) also validated the impact of the OMXC market in Denmark (-6.0 basis points) and the China News Agency on the Shanghai stock exchange (-9.5 basis points). This result is robust and statistically significant after extending the data set's time.

The results show the average effect of sentiment on DJIA returns and the ability of the approach to calculate statistically significant sentiment variables and have explanatory information about changes in financial returns. The rolling regression model is also estimated by the modelling part

of the approach to assessing the explanatory power of sentiment. Applying this method to Dow Jones Industrial Average returns and using negative sentiment proxies as independent variables, it is found that the sentiments' forecast power on returns were higher than the returns' forecast power on sentiments. News will often cover and reflect incidents that have taken place and affect the market. The information will already be included in the price, and the publication will trace the incident.

The case study in Chapter 4 also examines firm-specific text sentiment and firm-level stock returns. Vector autoregression analysis found negative sentiment and significant forecasts for some companies' return on the second day, mostly at the 1% level. One standard deviation increase in negative sentiment today will have a -5.7 to -16.1 basis points impact on the return of the stock on the second day. There is also evidence that negative sentiment on the second day would have the reverse causal effect of earnings. The difference between the negative sentiment scores today and yesterday has a stronger effect on today's returns (significant at 5%).

Using the rolling regression model, sentiment is thought to favour the explanatory power of the model before the market recession. First, the average impact of negative sentiment is compared in the rolling regression model with the expansion of the Dow Jones industrial average return more than the recession, but observing firm-level effects revealed the number of interpretable days of expansion (average 30 days/year) is one-third higher than the average effect of the depression period (average 19 days/year). Second, by comparing the periods of expansion and depression, one can see that the correlation between positive sentiment and negative sentiment is 20% on average during the expansion period. This also confirms that sentiment is more conducive to explain the model during expansion.

At the firm level, the most obvious relationship is the annual correlation. LOWESS regression results show that predictability tends to be concentrated in discrete periods, most likely in line with the important news story line for each company. Sentiment has less indirect repercussions, and the two effects do not often co-exist.

These results are in line with the literature (e.g. Tetlock (2007), Tetlock et al. (2008), Garcia (2012), etc.). Negative sentiment was found to have a direct negative impact on stock returns

and negative sentiment was more likely to explain stock returns during market expansion. These new contributions to the literature are as follows. First, the study recreates and extends the experimental results of Tetlock (2007) and validates the method's reliability in the Danish and Chinese markets. Second, the study generated reports of a firm-specific sentiment on news coverage of 23 large multinational companies, most of whom had 15 years of daily observation. This allows time series analysis to assess the correlation between several sentiment lags, returns, and volume of transactions. Most importantly, the study used non-linear regression analysis to visualise the relationship between negative sentiment and stock return residuals and used statistical methods to determine the discrete relationship between the benchmark index level and firm level.

5.2 Limitations and Future Work

One limitation is that whether the BoW method is still an updated method of extracting text sentiment in the future and whether the news text indicates real information in the financial market. This thesis extracts the sentiments and emotions of the text by calculating the frequency of appearance of the special vocabulary that is pre-determined in the dictionary. These vocabulary words were created by Stone et al. in the 1960s for the purpose of sociology and psychology research. They were used to create GI program and Harvard IV-4 dictionary. Whether these terms and phrases will continue to be interpretive and applicative over time is a consideration of the continued use of the GI program and the Harvard IV-4 dictionary in the sentiment analysis. The exploration of methods extended to machine learning will assist in the study of sentiment analysis. Machine learning can automatically analyse and obtain the rules from the data, and use the rules to predict the unknown data to access the data and estimate the performance of sentiment, which can improve the applicability of sentiment analysis.

Another factor that needs to be considered in the future is how to extract text sentiment at different levels, which determines whether or not in the news text the representation of market sentiment can be improved. The independence of news (relative to the market) is also a factor to

consider. This leads to the discussion of the second limitation. In the process of collecting data, it is found that news at the market level is often relatively straightforward to collect and sentiment extracted mostly reflect market sentiment. This is because the description of the relevant news in the market is the specific description, and there is nearly no duplicate description. However, when collecting firm-level news articles, the company's mentions in each article is examined and it is found that company news contained a large number of overlapping (see Table 4.4). This has resulted in a weakening of the independence of firm-level news. As a result, when studying the sentiment of a company, the data often carries the information of another company. How to separate and extract more pure and accurate firm-level sentiment becomes more interesting.

The third potential limitation is that the coverage of the news at the firm-level (largely) determines the value of the extracted sentiment. At the market level, the daily coverage of news is better (as mention of any of the companies included in the index can be regarded to the market), which directly leads to higher possibilities for obtaining sentiment data. But at the firm-level (where it is focused on the 23 companies), specific information and news are much harder to obtain. One reason is that the news coverage of an individual company is much lower than an index; secondly, news overlaps can lead to loss of information - a positive news article about a company and another positive news about a competitor of this company cancel each other out, and the extractable sentiment data will decrease. From a methodological point of view, it is difficult to try to extract data from a company and its competitors at the same time. Firm-level news reacts very quickly to sentiment. It will require particular targeted news. Company disclosures and financial statements (US 10-K forms, corporate financial statements, etc.) may well cover missing information. There is little overlap between companies in these financial reports.

Limited sources of sentiment extraction are also an extension of the direction worth considering. So far, this thesis only uses text as a source of sentiment. Data show that the limit of human typing speed is 216 English words per minute[72]. That is, if this research goes up to high-frequency tick data, one can only extract sentiment from a tiny sample of data (on the premise that the data provider is immediately available and the data is valid). If the range of data samples for extracting sentiment can be upgraded to voice and facial expression samples, the amount and

frequency of extractable data can be greatly increased. The process of sentiment analysis can be simultaneously developed as a large-scale test for real-time coverage of radio, television, online announcement, etc. Data sources include social media such as Facebook, Twitter and audio and video clips of YouTube and news portals.

Bibliography

[1] Khurshid Ahmad, Nicholas Daly, and Vanessa Liston. What is new? news media, general elections, sentiment, and named entities. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 80–88, 2011.

[2] Khurshid Ahmad, JingGuang Han, Elaine Hutson, Colm Kearney, and Sha Liu. Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance*, 2015.

[3] Raimund Alt, Ines Fortin, and Simon Weinberger. The monday effect revisited: An alternative testing approach. *Journal of Empirical Finance*, 18(3):447–460, 2011.

[4] Torben G Andersen and Tim Bollerslev. Deutsche mark–dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *The Journal of Finance*, 53(1):219–265, 1998.

[5] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Clara Vega. Micro effects of macro announcements: Real-time price discovery in foreign exchange. Technical report, National Bureau of Economic Research, 2002.

[6] Werner. Antweiler and Murray Z. Frank. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

[7] Robert A Ariel. A monthly effect in stock returns. *Journal of Financial Economics*, 18(1):161–174, 1987.

[8] Robert A Ariel. High stock returns before holidays: Existence and evidence on possible causes. *The Journal of Finance*, 45(5):1611–1626, 1990.

[9] Louis Bachelier. Theory of speculation. *The random character of stock market prices*, MIT Press, Cambridge, Mass. Blattberg 1018:17–78, 1900.

[10] Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.

[11] Pietro Balestra and Marc Nerlove. Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica: Journal of the Econometric Society*, pages 585–612, 1966.

[12] George A Barnard. New methods of quality control. *Journal of the Royal Statistical Society. Series A (General)*, 126(2):255–258, 1963.

[13] Söhnke M Bartram, Gregory Brown, and René M Stulz. Why are us stocks more volatile? *The Journal of Finance*, 67(4):1329–1370, 2012.

[14] Nittai K Bergman and Sugata Roychowdhury. Investor sentiment and corporate disclosure. *Journal of Accounting Research*, 46(5):1057–1083, 2008.

[15] Gene Birz and Sandip Dutta. Us macroeconomic news and international stock prices: Evidence from newspaper coverage. *Accounting and Finance Research*, 5(1):247, 2016.

[16] Fischer Black. Noise. *The Journal of Finance*, 41(3):528–543, 1986.

[17] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[18] John C Bogle. The index mutual fund: 40 years of growth, change, and challenge. *Financial Analysts Journal*, 72(1):9–13, 2016.

[19] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *The International AAAI Conference on Web and Social Media (ICWSM)*, 11:450–453, 2011.

[20] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[21] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[22] Kam Fong Chan, Robert G Bowman, and Christopher J Neely. Systematic cojumps, market component portfolios and scheduled macroeconomic announcements. *Journal of Empirical Finance*, 43:43–58, 2017.

[23] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[24] William S Cleveland. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35(1):54–54, 1981.

[25] Jason A Cook, Zeyan Zhao, and Khurshid Ahmad. Stylized facts of linguistic corpora: Exploring the lexical properties of affect in news. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 494–502. Springer, 2016.

[26] Frank Cross. The behavior of stock prices on fridays and mondays. *Financial Analysts Journal*, 29(6):67–69, 1973.

[27] Jesús Crespo Cuaresma, Gernot Doppelhofer, and Martin Feldkircher. The determinants of economic growth in european regions. *Regional Studies*, 48(1):44–67, 2014.

[28] David M Cutler, James M Poterba, and Lawrence H Summers. What moves stock prices? *The Journal of Portfolio Management*, 15(3):4–12, 1989.

[29] Zhi Da, Joseph Engelberg, and Pengjie Gao. The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies*, 28(1):1–32, 2014.

[30] Aswath Damodaran. The weekend effect in information releases: A study of earnings and dividend announcements. *The Review of Financial Studies*, 2(4):607–623, 1989.

[31] Sanjiv R Das and Mike Y Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[32] Joseph E Engelberg and Christopher A Parsons. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97, 2011.

[33] R Engle. Nobel lecture: Risk and volatility. *New York University*, 2003.

[34] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276, 1987.

[35] Eugene F Fama. Efficient capital markets: A review of theory and empirical work*. *The Journal of Finance*, 25(2):383–417, 1970.

[36] Andrey Feuerverger, Yu He, Shashi Khatri, et al. Statistical significance of the netflix challenge. *Statistical Science*, 27(2):202–231, 2012.

[37] Morris J Fields. Stock prices: a problem in verification. *The Journal of Business of the University of Chicago*, 4(4):415–418, 1931.

[38] Kenneth R French. Stock returns and the weekend effect. *Journal of Financial Economics*, 8(1):55–69, 1980.

[39] International Monetary Fund. World economic outlook database, 2017. https://www.imf.org/external/pubs/ft/weo/2017/02/weodata/index.aspx, 2018-03-31.

[40] D. Garcia. Sentiment during Recessions. *The Journal of Finance*, LXVIII(3):1267–1300, 2013.

[41] Clive W. J. Granger and Paul Newbold. *Forecasting Economic Time Series*. Academic Press: New York, 1977.

[42] Iain Hardie and Donald MacKenzie. Assembling an economic actor: the agencement of a hedge fund. *The Sociological Review*, 55(1):57–80, 2007.

[43] David Hirshleifer. Investor psychology and asset pricing. *The Journal of Finance*, 56(4):1533–1597, 2001.

[44] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

[45] Fang Jin, Nathan Self, Parang Saraf, Patrick Butler, Wei Wang, and Naren Ramakrishnan. Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1470–1473. ACM, 2013.

[46] Colm Kearney and Andrew J Patton. Multivariate garch modeling of exchange rate volatility transmission in the european monetary system. *Financial Review*, 35(1):29–48, 2000.

[47] Stephen 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.

[48] Stephen Kelly and Khurshid Ahmad. The impact of news media and affect in financial markets. In *Intelligent Data Engineering and Automated Learning–IDEAL 2015*, pages 535–540. Springer, 2015.

[49] Maurice Kendall. *Time Series*. Charles Griffin: London, 1976.

[50] Chan-Wung Kim and Jinwoo Park. Holiday effects and stock returns: Further evidence. *Journal of Financial and Quantitative Analysis*, 29(1):145–157, 1994.

[51] Peter Klibanoff, Owen Lamont, and Thierry A Wizman. Investor reaction to salient news in closed-end country funds. *The Journal of Finance*, 53(2):673–699, 1998.

[52] Gregory Koutmos and G Geoffrey Booth. Asymmetric volatility transmission in international stock markets. *Journal of International Money and Finance*, 14(6):747–762, 1995.

[53] Alok Kumar and Charles Lee. Retail investor sentiment and return comovements. *The Journal of Finance*, 61(5):2451–2486, 2006.

[54] Josef Lakonishok and Seymour Smidt. Are seasonal anomalies real? a ninety-year perspective. *The Review of Financial Studies*, 1(4):403–425, 1988.

[55] Hellmut Lehmann-Haupt. *Gutenberg and the Master of the Playing Cards*. Yale University Press:, 1966.

[56] Feng Li. The information content of forward-looking statements in corporate filingsa naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

[57] John Lintner. Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4):587–615, 1965.

[58] Tim. Loughran and Bill McDonald. When is a Liability not a Liability. *The Journal of Finance*, 66:35–65, 2011.

[59] Heather Lovell. Climate change, markets and standards: the case of financial accounting. *Economy and Society*, 43(2):260–284, 2014.

[60] Xing Lu, Jamshid Mehran, and Han Gao. Holiday trading in china: Before and during the financial crisis. *Journal of Applied Finance and Banking*, 6(2):117, 2016.

[61] Robert C Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470, 1974.

[62] G Mujtaba Mian and Srinivasan Sankaraguruswamy. Investor sentiment and stock market response to earnings news. *The Accounting Review*, 87(4):1357–1384, 2012.

[63] Franco Modigliani and Merton H Miller. The cost of capital, corporation finance and the theory of investment. *The American Economic Review*, 48(3):261–297, 1958.

[64] Jan Mossin. Equilibrium in a capital asset market. *Econometrica: Journal of the Econometric Society*, pages 768–783, 1966.

[65] Daniel Mügge and Bart Stellinga. The unstable core of global finance: Contingent valuation and governance of international accounting standards. *Regulation & Governance*, 9(1):47–62, 2015.

[66] Tommie Murphy, Stephen Kelly, and Khurshid Ahmad. Innovations in the crude oil market: Sentiment, exploration and production methods. *Exploration and Production Methods (April 14, 2015)*, 2015.

[67] W. K. Newey and K. D. West. A Simple, Positive Semi-Definite, Heteroscedastic and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55:703–709, 1987.

[68] OECD. Gross domestic product (expenditure approach) - per head, US $, current prices, current ppps, 2014. http://stats.oecd.org/index.aspx?queryid=558, 2014-05-25.

[69] National Bureau of Economic Research. US business cycle expansions and contractions, 2010. http://www.nber.org/cycles/cyclesmain.html, 2018-03-31.

[70] Joel Peress. The media and the diffusion of information in financial markets: Evidence from newspaper strikes. *The Journal of Finance*, 69(5):2007–2043, 2014.

[71] PD Praetz. A spectral analysis of australian share prices. *Australian Economic Papers*, 12(20):70–78, 1973.

[72] RATATYPE. Average typing speed infographic, 2018. https://www.ratatype.com/learn/average-typing-speed/, 2018-03-31.

[73] Richard Roll and Stephen A Ross. An empirical investigation of the arbitrage pricing theory. *The Journal of Finance*, 35(5):1073–1103, 1980.

[74] Michael S Rozeff and William R Kinney. Capital market seasonality: The case of stock returns. *Journal of Financial Economics*, 3(4):379–402, 1976.

[75] Mark Rubinstein. Rational markets: yes or no? the affirmative case. *Financial Analysts Journal*, 57(3):15–29, 2001.

[76] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[77] Robert P Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464, 2012.

[78] G William Schwert. Why does stock market volatility change over time? *The Journal of Finance*, 44(5):1115–1153, 1989.

[79] G William Schwert. Anomalies and market efficiency. *Handbook of the Economics of Finance*, 1:939–974, 2003.

[80] Mostafa Seif, Paul Docherty, and Abul Shamsuddin. Seasonal anomalies in advanced emerging stock markets. *The Quarterly Review of Economics and Finance*, 66:169–181, 2017.

[81] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964.

[82] Robert J Shiller, Stanley Fischer, and Benjamin M Friedman. Stock prices and social dynamics. *Brookings Papers on Economic Activity*, 1984(2):457–510, 1984.

[83] J Shiller Robert. Irrational exuberance. *Princeton, New Jersey, Princeton University*, 2000.

[84] Jeremy J Siegel. Equity risk premia, corporate profit forecasts, and investor sentiment around the stock crash of october 1987. *Journal of Business*, pages 557–570, 1992.

[85] Nate Silver. *The signal and the noise: the art and science of prediction.* Penguin UK, 2012.

[86] Christopher A. Sims. Macroeconomics and Reality. *Econometrica*, 48:1–48, 1980.

[87] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Olgilvie, and with associates. *The General Inquirer: A Computer Approach to Content Analysis.* The MIT Press, 1966.

[88] Ryan Sullivan, Allan Timmermann, and Halbert White. Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, 105(1):249–286, 2001.

[89] S Taylor. Modelling financial time series. *New York: Vlfiley*, 1986.

[90] S. J. Taylor. *Asset Price Dynamics, Volatility, and Prediction.* Princeton University Press, 2011.

[91] Stephen J Taylor. Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2):183–204, 1994.

[92] Paul C. Tetlock. Giving Content to Investor Sentiment: The Role of Media in the Stockmarket. *The Journal of Finance*, 62(3):1139–1168, 2007.

[93] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[94] Jessica Toonkel. Blackrock betting big data can help revive its active equity funds, 2015. https://www.reuters.com/article/us-blackrock-bigdata-equity-analysis/blackrock-betting-big-data-can-help-revive-its-active-equity-funds-idUSKCN0QB0B520150806, 2018-03-31.

[95] Tsien Tsuen-Hsuin and Joseph Needham. *Science and Civilisation in China: Chemistry and Chemical Technology. Paper and Printing.* Cambridge University Press, 1985.

[96] Amos Tversky and Daniel Kahneman. Rational choice and the framing of decisions. *Journal of Business*, pages S251–S278, 1986.

[97] Hal R Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

[98] Robert E Whaley. The investor fear gauge. *The Journal of Portfolio Management*, 26(3):12–17, 2000.

[99] Jialin Yu. Disagreement and return predictability of stock portfolios. *Journal of Financial Economics*, 99(1):162–183, 2011.

[100] Yang Yu, Wenjing Duan, and Qing Cao. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4):919–926, 2013.

[101] Tian Yuan and Rakesh Gupta. Chinese lunar new year effect in asian stock markets, 1999–2012. *The Quarterly Review of Economics and Finance*, 54(4):529–537, 2014.

[102] Udny G. Yule. *An Introduction to the Theory of Statistics*. Charles Griffin: London, 1922.

[103] Zeyan Zhao and Khurshid Ahmad. Qualitative and quantitative sentiment proxies: Interaction between markets. In *Intelligent Data Engineering and Automated Learning–IDEAL 2015*, pages 466–474. Springer, 2015a.

[104] Zeyan Zhao and Khurshid Ahmad. A computational account of investor behaviour in chinese and us market. *International Journal of Economic Behavior and Organization*, 3(6):78–84, 2015b.

[105] Zeyan Zhao, Stephen Kelly, and Khurshid Ahmad. Finding sentiment in noise: Non-linear relationships between sentiment and financial markets. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 580–591. Springer, 2017.

[106] William T Ziemba. Japanese security market regularities: Monthly, turn-of-the-month and year, holiday and golden week effects. *Japan and the World Economy*, 3(2):119–146, 1991.