INFLUENCE OF AGGREGATE NEWS SENTIMENT ON INDIAN STOCK MARKET IN THE SHORT RUN

A Thesis submitted in partial fulfillment for the Degree of

DOCTOR OF PHILOSOPHY

in the

Goa Business School, Goa University

by

SUSHANT GANPAT CHARI

under the guidance of

DR. PURVA HEGDE DESAI

Professor, Goa Business School

Goa University

Taleigao - Goa

(May - 2022)

DECLARATION

I, Mr. Sushant Ganpat Chari, hereby declare that this thesis titled **"Influence of aggregate news sentiment on Indian Stock Market in the short run"** represents work which has been carried out by me and that it has not been submitted, either in part or full, to any other University or Institution for the award of any research degree.

Sushant Ganpat Chari

Place: Goa University
Date:

CERTIFICATE

This is to certify that the Ph.D. thesis titled **"Influence of aggregate news sentiment on Indian Stock Market in the short run"** is an original work carried out by Mr. Sushant Ganpat Chari under my guidance, at the Management Discipline, Goa Business School, Goa University. This thesis or a part thereof, has not formed the basis for the award of any Degree, Diploma, Title or Recognition before.

Dr. Purva Hegde Desai
Professor,
Management Discipline,
Goa Business School,
Goa University.

Place: Goa University
Date:

DEDICATION

This thesis is dedicated to my parents Smt. Radhabai G. Chari and Late. Shri. Ganpat S. Chari who instilled in me the virtues of hard work, perseverance, commitment and patience.
Thank you for your blessings and effort.

ACKNOWLEDGEMENT

As I moved towards completion of my thesis, I pondered over the journey of my research. This journey began with a hope of exploring the stock market and its dynamic behavior. In this journey I came across situations and issues that required not only my mental strength to withstand criticisms and denials but also the motivation, guidance and support from people around me. Before I thank them, I take this opportunity to express my deep sense of gratitude to God for showering blessings on me throughout my life, making me believe that I can achieve success in this research journey and standing by me in difficult circumstances.

I feel myself proud to have conducted my research under the guidance of Dr. Purva G. Hegde Desai, Professor, Management discipline, Goa Business School, Goa University who kept faith in my ideas throughout the duration of this research period. Her immaculate guidance, motivation, commitment and knowledge encouraged me to always strive for quality in my research work. I am very grateful to her for shaping my research abilities and inspiring me to conduct this research.

I am deeply grateful to Dr. Nilesh Borde, Professor, Management discipline, Goa Business School, Goa University whose invaluable inputs and suggestions enlightened me to think about my research from different perspectives. I still remember, the day when he probed my research proposal by taking the role of a devil's advocate, cautioning me on some of my ideas and also praising me on others. I am deeply indebted to him for the amount of time and energy he spent to fine tune my research and interact with me on numerous occasions.

I am also thankful to the subject experts on Departmental Research Committee (DRC) viz. Dr. Pranab Mukhopadhyay, Professor, Economics Discipline, Goa Business School, Goa University and Dr. Harip Khanapuri, Associate Professor, Goa Business School, Goa University for their constructive criticisms as well as providing direction to my research.

I appreciate the contribution of other faculties viz. Dr. Nandkumar Mekoth, Dr. M. S. Dayanand and Dr. Nirmala R. from Management discipline, Goa Business School who provided clarity to certain aspects of my research by giving suggestions and also questioning me on those areas of my research which had ambiguity. This helped me understand and improve my research.

I would also like to thank all my co-researchers in Management discipline at Goa business school who provided me their moral support and encouragement for presentations or otherwise. At the same time, I would also like to appreciate the support I received from the office staff of Goa business school, especially Mr. Amrut Shet and Mr. Uday Borkar. Similarly, I would also like to thank the staff of Computer Centre, Library and Administrative department for their support at various stages of my research study.

I would like to convey my special thanks and appreciation to my colleagues from Saraswat Vidyalaya's Sridora Caculo College of Commerce and Management Studies, Khorlim, Mapusa, Goa including the Principal Prof. Santosh B. Patkar, Dr. Abhishek Karmali, Mr. Rajeev Narvekar, Dr. Harsha Talaulikar and Ms. Pooja Lawande Karmali for supporting and encouraging me to pursue research.

I would not have reached here to become a researcher without the efforts, motivation, commitment and belief of my family members during various stages of my life. I am indebted to my parents Smt. Radhabai G. Chari and Late Shri. Ganpat S. Chari. I am also thankful to my sister Ms. Ashalata Chari, my brothers Mr. Pritam Chari and Mr. Manojkumar Chari, my wife Ms. Dhanashri Chari and my sister-in-law Ms. Nisha M. Chari for their efforts and motivating me to pursue my academic goals.

Abstract

INFLUENCE OF AGGREGATE NEWS SENTIMENT ON INDIAN STOCK MARKET IN THE SHORT RUN

By: Sushant G. Chari

Research Guide: Dr. Purva G. Hegde Desai, Professor, Management Discipline, Goa Business School, Goa University

In today's world of electronic communication, information access has become easy and we can come to know about any event happening around the world at a lightning speed. People also provide their opinions and reactions about the event via social media in quick time. Often we make use of this information in taking decisions, be it about our travel plan, business strategy or investment in an asset or any other activity for which we need to take a decision. For instance, if there is some information about terrorist attack in some area where we had plans to go on a tour; we are likely to change our plan and select a different destination. So is the case with investment decisions. If there is some news about a company indulging in corruption, most of us will change our decision and buy a different stock than the one for which we received negative news. Investors are all human beings making decisions in the stock market to buy, hold or sell stocks with a mind that is affected by the information that they receive subject to their belief systems, circumstances and the prevailing market environment. While investor's belief system, circumstances and the prevailing market environment is relatively static during a short period of time, the sentiment conveyed in the information received by the investors is dynamic and can affect those static factors either positively or negatively. This makes it very interesting to find out what kind of information really affects investors; particularly the case where most of them think alike and participate in the stock market resulting in extreme movement. On a given day, an investor or a trader must be accessing variety of news related to market events as well as other types of events. So, surely he/she will not be accessing a single piece of news in isolation; rather, it would be a mix of several news he/she has an interest on.

Events like war, general elections, budget, etc. have extensive coverage on media. A large number of news articles, columns, opinions, etc. are published during such events; sometimes overpowering the news coverage of other smaller events. Because of this, sometimes public mood is polarised in a particular direction. It should also be noted that public mood is a result of expectations built by the information they receive before, during and after the event. Sometimes these expectations may exceed the reality which are reflected in their behavior. Especially, investors who tend to trade on information sometimes have expectations which do not meet the reality and hence their participation in the stock market tend to deviate market prices away from the real value (intrinsic value). The case of extreme stock market movement requires investigation

with reference to the influence of news sentiment generated from news related to business, economy, international and political events.

At the backdrop of the above discussion, the focus of this thesis is on addressing the following concerns which have not been explored enough in the extant literature:

1. Influence of aggregate news sentiment on returns of stock market: Past studies on company-specific news sentiment only consider news sentiment from news released on the company without due attention towards other news having a potential to influence investors in their decision making. This doesn't depict the realistic situation an investor goes through on any given day as they receive different news with a potential to affect their decision making. The focus of this thesis has been on examining the influence of aggregate news sentiment generated as a result of news from business, economy, politics and international categories on returns of stock market and sectoral indices specifically when there is extreme movement in stock market.

2. Influence of news sentiment from different categories of news on returns of stock market: A study by Cutler et al. (1989) showed that movement in stock prices can be explained using variables related to the company as well as variables related to macroeconomic, political and international events. When information about these events is released, investors have certain expectations from it but cannot estimate the value of this information correctly because of which the market prices differs from the actual price. This difference can be attributed to the expectations formed from the information investors receive; it represents feelings of investors about the event referred in the information. Prior studies have given evidence for the influence of company-specific news sentiment on their related stock's variables like returns, volatility, trading volume, etc. However, news sentiment related to economy, politics and international events have not been studied enough for their influence on the stock market.

3. Influence of aggregate news sentiment and category-wise news sentiment on returns of sectoral indices: Influence of company-specific news sentiment on its related stock variables is well documented in the extant literature. Since, the focus of this study is on news sentiment related to politics, international and economy apart from business related news sentiment, the influence is expected to be on large number of stocks rather than a specific stock. If this is the case then it makes sense to examine portfolios of stocks and understand how aggregate news sentiment and category-wise news sentiment influences them. In this thesis, such study has been conducted to understand how aggregate news sentiment and category-wise news sentiment influences returns of sectoral indices which are portfolios created by NSE based on important industry sectors in India.

In short, this study has been conducted to find out the influence of aggregate news sentiment derived from business, economy, politics and international news categories on returns of stock market and select sectoral indices during extreme market movement. Keeping noise trader theory as the background theory for this research, data points corresponding to extreme movement in Nifty50 index are filtered by setting a cutoff of (+/- 2.58)σ where σ is the daily standard deviation estimated using GARCH(1,1) model. By setting an observation window of five trading days around the day of extreme market returns, returns of all trading days from all the observation windows are obtained. News published related to business, economy, politics and international news are obtained from the archives of "economictimes.indiatimes.com" for all days corresponding to all observation windows. After preprocessing steps, sentiment of each news article is obtained and then aggregated on daily basis as well as according to categories on daily basis for all days in the observation window. Using linear regression, relationship between aggregate news sentiment and returns of Nifty50 index is examined to check influence of aggregate news sentiment on returns of Nifty50 index. Similarly, using multiple regression, influence of news sentiment of business, economy, politics and international news categories on returns of Nifty50 index is examined. Also, spillover effect of aggregate news sentiment of previous two days on the returns of Nifty50 index on the day of extreme market movement is examined. Further, influence of aggregate news sentiment on returns of select sectoral indices is examined during extreme movement in Nifty50 index and also influence of news sentiment of business, economy, politics and international categories on returns of select sectoral indices are examined.

The results obtained in this thesis suggest that aggregate news sentiment influences returns of stock market (Nifty50 index) on the day of extreme market movement. Also, relatively weaker influence of aggregate news sentiment on returns of stock market is found a day prior to the day of extreme market movement. News sentiment from business, economy, politics and international news categories are found to be having a significant influence on returns of stock market on the day of extreme market movement. Here business and politics news categories are found to be significant predictors of the returns of stock market. Evidence for spillover effect of aggregate news sentiment of previous day of extreme market movement on returns of stock market on the day of extreme market movement is found to be significant at 10% significant level. When aggregated news sentiment is examined for its influence on returns of sectoral indices, it is found to be having a significant influence on the returns of select sectoral indices on the day of extreme market movement. News sentiment from business, economy, politics and international categories is found to influence returns of sectoral indices on the day of extreme market movement. Similar to stock market, business and politics news sentiment are found to be significant predictors of returns of sectoral indices.

The study conducted in this thesis is based on the premise that noise traders are active when prices deviate from the intrinsic value and are affected by sentiment induced by the news that they receive about events having potential to influence stock market. The major contribution of this thesis has been in establishing the influence of aggregate news sentiment and category-wise news sentiment on returns of stock market thereby associating news sentiment with the sentiment discussed in the noise trader theory. Second, similar to studies that found influence of company-specific news sentiment on returns of its related stock, in this thesis news sentiment from business and politics news categories have also been found to have significant influence on returns of portfolios of stocks like Nifty50 index and select sectoral indices. Third, this thesis also provides an evidence for short term influence of aggregate and category-wise news sentiment on returns of stock market and select sectoral indices which has a major implications for practitioners in the market suggesting them against taking any bold decisions on the day of extreme market movement as it would be disastrous in the future because the influence is for a short term.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research conducted within the scope of this thesis. To begin with, section 1.1 provides background of the study. This is followed by section 1.2 that presents the overview of Global and Indian stock market. Thereafter sections 1.3, **??**, 1.4, and 1.5 presents statement of the problem, objectives of the study, motivation and significance of the study, and scope of the study respectively. This chapter concludes by giving an overview of the methodology in section 1.6 followed by organization of chapters of this thesis in section 1.7.

1.1 Background

In today's digital age, information is distributed at a lightning pace. News, in particular being a publicly shared information, carries a feeling about the events it covers across the mass audience. If the event on which news is published is such that it affects the macro-environment in the country, it may create a feeling of pessimism or optimism in the audience. Investors in particular who invest in the stock market often find themselves in the tides of such feelings which generate a tendency in them to trade on signals deciphered from the news they receive. In today's times, ease of access to such news has increased the participation of retail investors who feel the pulse of the movement on every tick in the stock price. Along with the capabilities of electronic media, advancements in technology used to carry out transactions in the stock market has also resulted in higher participation of investors as they can now transact irrespective of the physical location where they are present. Thus, electronic mode of communication has provided a convenient access to information.

If we ask ourselves a question- what is the use of the new information that we get through media and what do we do with it? Definitely many of us will say it updates our knowledge and we use it for taking decisions in our day to day life. But how does this updation takes place? Knowledge can be viewed in the form of certain beliefs about a subject or an entity. When we get new information, it either re-enforces our existing beliefs or suppresses them depending on the strength of existing beliefs and the strength of feelings conveyed in the new information received. In either case, our knowledge gets updated through our belief system. If we take the case of market participants who are irrational in their behavior, content of new information may be such that his/her already held beliefs about a stock may change or may get further re-enforced with new information. Moreover, sometimes the information received may develop a sense of optimism or pessimism that can trigger investors or traders to participate in the market.

Ever since the stock markets came into existence, investigating the relationship between information and the stock market movement has been a strand of research many researchers found interesting. While the mode of access to information and mode of transactions in stock markets evolved over a period of time, so has been the research conducted on it; every researcher adding new dimensions and perspectives to this relationship.

A few decades back, we had to rely on print media to provide us with new information, which was quite slow in comparison to today's electronic media. Although, radio and television were existent in the recent past and are available even today, one needs to be in front of them all the time to get updated. This is where online media has a great advantage in terms of its reach. It can be accessed at any time and whichever location provided network connectivity is available. Unlike radio or television, some online media sources provide news for reference even after several days from its release. Another advantage of online sources is that they are available on the fly on mobile or laptop and being digital, allows one to store and process it further. Today we see a large number of online sources that not only provide information about events happening around the world but allows one to express their opinions and feelings.

As mentioned earlier, ever since researchers began their study on stock markets, one of the key themes that found a place in their research was how stock markets react to new information. One of the early theories that provided foundations for understanding this was the Efficient Market Hypothesis (EMH) proposed by Fama, 1965. Thereafter, several other researchers, especially from behavioral finance like Hong and Stein (1997), Daniel et al. (1998), Barberis et al. (1998), Barber and Odean (2008), Fenzl and Pelzmann (2012), etc. provided theoretical explanations and evidences for stock market movement. While classical finance theories like EMH did not take into consideration the role of investors, behavioral finance researchers felt the need to include the psychological and sociological biases of investors. Also, various event studies as summarized in Mcwilliams and Siegel (1997) showed how company-specific events affect stock returns. Jareño and Negrut (2016) and Humpe and Macmillan (2009) showed how macroeconomic variables like consumer price index (CPI), gross domestic product (GDP), unemployment figures, interest rates, Industrial Production Index, price of crude oil, etc. affect stock markets. One thing that was common across these studies was the use of quantitative data. However, in the real world, quantitative data is not the only data available in an isolated manner; there is always a context to it. For instance, when quantitative data like earnings of a company are released, there is always a qualitative data in the form of narrative or guidance that is also provided by the Company. So, along with quantitative data, there is also a qualitative data available that gives context to that data and is often subjective in nature. Even when the foundation theories on the stock market were proposed, qualitative data was present, but it was available mostly in the printed form. Because of the lack of technology to acquire and process qualitative data, only quantitative data was used in most of the research, and qualitative data was neglected. However, things are very different now than they used to be. Even though today print media exists, there is a large volume of qualitative data generated in electronic form which can be acquired, processed, stored, and disseminated with ease. Day by day, along with quantitative data, a massive amount of qualitative data is being generated in the form of guidelines, narratives, press releases, news, etc. In addition to this, qualitative data is also available on social media platforms expressed as opinions, views, reviews, etc. of various stakeholders. The growing popularity and the sheer volume of news and social media data generated in the world today show the extent of consumption of qualitative data among readers. This is a huge resource that remained unexplored for a long period by researchers. However, in recent times researchers and corporates have started processing it by using a text processing technique called “Sentiment Analysis". Along with sentiment analysis, other text processing techniques like web mining, natural language processing, etc. are being employed to extract and process text available on news and social media platforms.

W. Zhang and Skiena (2010), Ferguson et al. (2011), Ranco et al. (2015) and many others have found sentiment analysis of news, blogs, twits, etc. useful to predict and design investment and trading strategies. Since news media is one of the most important carriers of live information, the general public including investors accesses it. Barber and Odean (2008) found that along with stock’s abnormal daily trading volume and its (previous) 1-day return, news on a firm tends to grab individual investor’s attention for stock picking. So, if the news carries information that creates a pessimistic or optimistic tone across a large community of investors about a stock or entire market, it may result in a significant movement in that stock or market.

One of the common methodologies in many research studies on sentiment derived from text involved a process of obtaining text content in the form of news, twits, blogs, Google search queries, etc. on company-specific events like earnings announcement, dividend declaration, mergers, and acquisition, etc. and analyzing it to examine if it influences certain market variables. It is true that the content used in these studies has a direct relationship with the Company’s performance and hence showed a relationship with related stock variables like returns, volatility, etc. However, if we follow events like elections, trade wars between countries, economic policy decisions, Government decisions in the parliament, etc. which are not company-specific but seems to move stock markets drastically. There is not enough evidence of research conducted either using sentiment analysis of content released during such events or otherwise. Interestingly, one of the significant features of such events is that there is a lot of buzz around them on news and social media platforms which gives rise to a feeling of pessimism or optimism in the market. The difficulty however is that there are too many such diverse events that happen together on a given day and hence the influence created by a particular event cannot be isolated and measured. Also, some of these events being unique, make it difficult to get sufficient data points to statistically prove a cause-effect relationship. One of the ways to conduct research under these constraints is to classify news into categories and then find out which news category contributed towards the stock market movement. Since news is published concerning a particular event, one may create news categories like politics, economy, international, sports, technology, etc., find news sentiment of each category, and then examine if they contribute towards movement in the stock market. Also, their sentiment may be combined to obtain aggregate news sentiment on a given day and examined to see if it influences the stock market. Though this method creates a rosy picture, aggregate news sentiment may not be able to influence the stock market every day as the strength of sentiment in the news may not be enough to overpower the fundamentals. In such circumstances, an ideal way to go ahead with such research is to find out instances when the market has deviated from the fundamentals.

The model of noise trader risk by De Long et al. (2003); also known as Noise Trader Theory posits that extreme deviations in the market is because of noise traders' sentiment and the limits on the arbitragers. With the premise that noise traders' beliefs get updated because of news about various events they receive, it is expected that influence of aggregate news sentiment will be stronger during extreme market movement due to higher probability of noise traders' participation who feel confident about their sentiment when stock market shows movement in the direction of their sentiment. Since this sentiment is not exclusively stock specific, it is expected that it would affect a wide variety of stocks. So, in such a study examining a portfolio of stocks is more relevant than the individual stocks.

In this thesis, the influence of aggregate news sentiment on Nifty50 index returns is examined. Trading dates with extreme Nifty50 index returns are filtered by setting a 2.58 standard deviation cut-off over 12 years. Keeping an observation window of five days comprising of two trading days before, the day of extreme market movement and two trading days after the day of extreme returns, news sentiment of archived news articles related to four news categories viz. business, economy, international, and politics is obtained from online news portal "economictimes.indiatimes.com". After finding news sentiment of each news article, aggregate news sentiment and category-wise news sentiment is obtained for each day in the observation window. Using linear regression, the influence of aggregate news sentiment on Nifty50 index returns for all five days in the window of observation is compared. Similarly, using multiple regression, the influence of news sentiment of each news category on Nifty50 index returns for all five days is compared. This analysis is then repeated by replacing the Nifty50 index with select sectoral indices during extreme movement in stock market (Nifty50 index).

1.2 Overview of Global and Indian Stock Market

Equity market or stock market is a market in which shares of publicly held companies are issued and traded either through exchanges or over-the-counter markets. Stock market enables the companies to get access to capital in exchange for a certain portion of ownership in the company. In order to provide services for stock brokers and traders to buy or sell stocks, bonds, and other securities, organization called a “stock exchange” is established. Stock markets today have become one of the most important constituents of economy and is known to be the barometer of country’s economy.

1.2.1 Global stock market

The history of stock markets dates back to 1602, when the Dutch East India Company officially became the worlds first publicly traded company to release shares of the company on the Amsterdam Stock Exchange. Today, almost every country in the world has its own stock market. Five most powerful stock exchanges in the world as per their market capitalization as on 19^{th} June 2020 are New York Stock Exchange (Market cap.:$19.3 trillion), NASDAQ (Market cap.: $13.3 Trillion), Tokyo Stock Exchange (Market cap.: $5.7 Trillion), Shanghai Stock Exchange (Market cap.: $4.9 Trillion) and Hong Kong stock Exchange (Market cap.: $4.4 Trilllion) (Source: www.businessinsider.in). Earlier the trading in stock markets happened using an outcry system that used to match buyers and sellers through the use of verbal bids and offer prices. However, today most of the stock exchanges around the world use electronic trading system. Electronic trading system performs order entry and forwarding, matching of buy and sell orders, and price determination using computer.

While the global stock markets are known to be powerful and has prospered many listed companies and the market participants, there have been instances in the past where markets crashed by a significant amount causing heavy losses to them. Some of the well-known crashes in the global stock market are listed below:

- Black Thursday or Terrible Thursday of 1929
- Stock Market Crash of 1973-1974
- Black Monday of 1987
- Dot-com Bubble of 2000
- Stock Market Crash of 2008

1.2.2 Indian stock Market

In India, Bombay Stock Exchange (BSE) was established in 1875 and is considered to be the Asia’s first stock exchange and also known as the fastest Stock Exchange in the world with a speed of 6 micro seconds (BSE, 2016). In 2017, BSE became the 1^{st} listed stock exchange of India. Currently, it provides an efficient and transparent market for trading various types of financial instruments such as equity, currencies, debt instruments, derivatives, mutual funds, etc. S&P BSE SENSEX is India's most widely tracked stock market benchmark index and is traded internationally on the EUREX and other leading exchanges of the BRICS nations (Brazil, Russia, China and South Africa).

The major transformation to the stock market and the overall financial system in India happened post liberalization in 1991 when the Indian Government introduced the economic policy to implement structural reforms to boost economy. During this period, a regulatory authority named as SEBI (Security Exchange Board of India) was introduced to monitor the activities of stock exchanges. Also National Stock Exchange (NSE) established in 1992 was recognised as a stock exchange by SEBI in April 1993 and commenced its operations in 1994.

National Stock Exchange: Background

NSE started trading in the equities segment on November 3, 1994. It started electronic screen-based trading in 1994, derivatives trading (in the form of index futures) and internet trading in 2000. It has a fully-integrated business model delivering services such as exchange listings, trading services, clearing and settlement services, indices, market data feeds, technology solutions and financial education offerings. It also administers compliance by trading and clearing members and listed companies with the rules and regulations of the exchange.

With rapid strides, today it has become the largest stock exchange in India in terms of volumes transacted. Also, as per World Federation of Exchange (WFE) Report - 2019, it is ranked as the third largest stock exchange globally in terms of number of equity trades. Within a short period of time, equity segment has shown a massive growth with average daily turnover increasing from Rs.17 crores during 1994-95 to around Rs. 32,475 crores as of December 2019. At the same time, NSE has listed over 1900 securities with market capitalization of over Rs 154.32 lakh Crores (NSE, 2015). NSE uses fully automated trading system called "National Exchange for Automated Trading" (NEAT) providing for order driven market.

Products of National Stock Exchange

NSE's products can be categorized into 3 asset classes for trading:

- Capital market for the listing and trading of equities: This asset class comprises of Equity and equity-linked products available for trading in the cash market such as stocks, IDRs, ETFs[1] (including NIFTY indices), units of closed-ended mutual fund schemes, and SME's listed on EMERGE.

[1]ETF: Exchange Traded Fund

- Fixed income securities and Debt products include Negotiated Trade Reporting in Government securities, Corporate Bonds, Sovereign Gold Bonds and other debt securities traded on multiple platforms

- Derivatives segment: This caters to derivative contracts on Equity, Indices, Currency, Interest Rates and Commodities.

Stock Indices of National Stock Exchange

National stock exchange has a subsidiary called NSE Indices Limited setup in May 1998 that provides a variety of indices and index related services and products for the Indian capital markets. Formerly it was known as India Index Services & Products Limited (IISL). As of March 31, 2019, it owns and manages a portfolio of 162 indices under NIFTY brand. NIFTY indices has served as the benchmark index for 53 ETFs listed in India and 9 ETFs listed abroad as of March 31, 2019. Also, derivatives benchmarked to NIFTY indices are traded on the Singapore Exchange, and the Taiwan Futures Exchange.

Categories of Stock indices

There are four categories of stock indices traded at NSE as given below:

1. Broad based Indices: These indices consist of the large, liquid stocks listed on the NSE and serve as a benchmark for measuring the performance of the stocks or portfolios such as mutual fund investments. Some examples include NIFTY 50, NIFTY NEXT 50, NIFTY 100, NIFTY 200, etc.

2. Sectoral Indices: These indices are created based on the major industrial sectors. In total there are 14 sectoral indices: NIFTY Auto Index, NIFTYBank Index, NIFTY Financial Services Index, NIFTY Financial Services 25/50 Index, NIFTY FMCG Index, NIFTY IT Index, NIFTY Media Index, NIFTY Metal Index, NIFTY Pharma Index, NIFTY Private Bank Index, NIFTY PSU Bank Index, NIFTY Realty Index, NIFTY Consumer Durables Index and NIFTY Oil and Gas Index.

3. Thematic indices: These indices are based on certain themes that reflect the performance of companies belonging to a particular corporate group (e.g. NIFTY Aditya Birla group, NIFTY Mahindra Group, NIFTY Tata group, etc.), diversified portfolio of companies representing some specific sectors (E.g. NIFTY INFRASTRUCTURE, NIFTY Energy Index, NIFTY Services Sector Index, etc.), companies within NIFTY 100 index, based on Environmental, Social and Governance (ESG) scores (e.g. NIFTY100 ESG Index, NIFTY100 Enhanced ESG Index, etc.), etc.

4. Strategy indices: These indices are designed on the basis of quantitative models or investment strategies to provide a single value for the aggregate performance of a number of companies (NIFTY High Beta 50 Index, NIFTY Alpha 50 index, NIFTY Dividend Opportunities 50 Index, etc.)

The Nifty50 index

The NIFTY 50 is a well diversified 50 stock index owned and managed by NSE Indices Limited and represents listed companies from important sectors of the economy. As on March 29, 2019, it represents about 66.8% of the free float market capitalization of the stocks listed on NSE. It is used for benchmarking fund portfolios, index-based derivatives and index funds. It is computed using a float-adjusted, market capitalization weighted methodology, wherein the level of the index reflects the total market value of all the stocks in the index relative to a particular base period.The base period selected here is the closing prices on the day of completion of one year of operations of NSE's Capital Market Segment (November 3, 1995); base value of the index being 1000 and a base capital of Rs.2.06 trillion. The review of NIFTY 50 happens semi-annually based on data for six months ending January and July. If there is any replacement of the stock(s) is to be done, then it is generally implemented from the first working day after F&O expiry of March and September and market participants are given four-week prior notice about this change. Additional details about Nifty50 index is provided in the Table B.1 and Table B.2 of Appendix B.

NSE sectoral indices

Over the years, NSE has designed a number of indices under its sectoral indices theme to reflect the collective performance of various listed companies belonging to a particular industrial sector. A brief idea about the various sectoral indices considered in this thesis is given below:

1. Nifty Auto Index: This index is designed to reflect the behavior and performance of the stocks of listed companies belonging to Automobiles sector which manufactures cars, motorcycles, heavy vehicles, auto ancillaries, tyres, etc. It consists of a maximum of 15 stocks. It was launched on July 12, 2011 by taking January 1, 2004 as the base date and with a base value of 1000. Additional details about Nifty Auto index is provided in the Table B.3 of Appendix B

2. Nifty Bank: This index is designed to reflect the behavior and performance of the most liquid and large capitalized Indian banks listed on NSE. It consists of a maximum of 12 stocks. It was launched on September 15, 2003 by taking January 1, 2000 as the base date and a base value of 1000.Additional details about Nifty Auto index is provided in the Table B.4 of Appendix B.

3. NIFTY FMCG Index: This index is designed to reflect the behavior and performance of Indian companies listed on NSE from the Fast Moving Consumer Goods (FMCG) sector that deals with goods and products which are non-durable, mass consumption products and available off the shelf. It consists of a maximum of 15 stocks. It was launched on September 22, 1999 by taking January 01, 1996 as the base date and a base value of 1000. Additional details about Nifty FMCG index is provided in the Table B.5 of Appendix B.

4. Nifty Energy index: This index is designed to reflect the behavior and performance of a diversified portfolio of companies representing the commodities segment which includes sectors such as Petroleum, Gas and Power, etc. It comprises of 10 companies listed on National Stock Exchange of India (NSE). It was launched on July 01, 2005 by taking January 01, 2001 as the base date and a base value of 1000. Additional details about Nifty Energy index is provided in the Table B.6 of Appendix B

5. NIFTY IT index: This index is designed to reflect the behavior and performance of Indian IT companies listed on NSE. It consists of a maximum of 10 stocks. Computed using free float market capitalization method with a base date of Jan 1, 1996 indexed to a base value of 1000, it was later revised to a base value of 100 with effect from May 28, 2004.Additional details about Nifty IT index is provided in the Table B.7 of Appendix B.

6. Nifty Media Index: This index is designed to reflect the behavior and performance of the Media & Entertainment sector including companies involved in printing and publishing in India. It consists of a maximum of 10 stocks. It was launched on July 19, 2011 by taking December 30, 2005 as the base date and a base value of 1000. Additional details about Nifty Media index is provided in the Table B.8 of Appendix B.

7. Nifty Pharma Index: This index is designed to reflect the behavior and performance of the pharmaceutical sector in India. It consists of a maximum of 10 stocks. It was launched on July 01, 2005 by taking January 01, 2001 as the base date and a base value of 1000. Additional details about Nifty Pharma index is provided in the Table B.9 of Appendix B.

8. Nifty Private Bank Index: This index is designed to reflect the behavior and performance of the banks from private sector. It consists of a maximum of 10 stocks. It was launched on January 05, 2016 by taking April 01, 2005 as the base date and a base value of 1000. Additional details about Nifty Private Bank index is provided in the Table B.10 of Appendix B.

9. Nifty PSU Bank Index: This index is designed to reflect the behavior and performance of the public sector banks. It consists of a maximum of 12 stocks. It was launched on August 30, 2007 by taking January 01, 2004 as the base date and a base value of 1000. Additional details about Nifty PSU Bank index is provided in the Table B.11 of Appendix B.

10. NIFTY Realty Index: This index is designed to reflect the performance of real estate companies listed on NSE and are involved in the construction of residential and commercial properties. It consists of a maximum of 10 stocks. It was launched on August 30, 2007 by taking December 29, 2006 as the base date and a base value of 1000. Additional details about Nifty Realty index is provided in the Table B.12 of Appendix B.

Over a period of time, the Indian stock market has grown tremendously in terms of the technology used for conducting transactions, the number of stocks listed on stock exchanges, market capitalization, trading volumes and turnover on stock exchanges, and more importantly the investor base.

1.3 Statement of the Problem

This thesis seeks to study the relationship between the aggregate news sentiment derived from news belonging to business, economy, politics and international categories and the returns of stock market during extreme movement in stock market. The problems that this research investigates can be stated as follows:

1. Does aggregate news sentiment derived from business, economy, politics and international categories has any relationship with returns of stock market during extreme movement in stock market?

2. Is there any influence of news sentiment of business, economy, politics and international news categories on the returns of stock market on the day of extreme stock market movement?

3. Is there any spillover effect of aggregate news sentiment of the previous two days on the returns of stock market during extreme movement in stock market?

4. Does aggregate news sentiment derived from business, economy, politics and international categories has any relationship with returns of sectoral indices during extreme movement in stock market?

5. Is there any influence of news sentiment of business, economy, politics and international news categories on the returns of sectoral indices during extreme movement in stock market?

1.4 Motivation and Significance of the study

1.4.1 Motivation:

In recent years, it is observed that stock markets around the globe have seen high volatility in the prices in the short run. If one keeps track of the stock market on a daily basis, he/she will be able to correlate some of these turbulences to some company announcement, political disturbance, economic indicator or macroeconomic policy decision, global geo-political tension, etc. For example, in Indian stock market telecom and mining companies showed a southward movement for news reported by media on their involvement in 2G spectrum and mining scam respectively. Another example is the Indian stock market's upward movement during the period around 2014 general election with the expectation of BJP (Bhartiya Janata Party) coming to power. A more recent short-term decline in the global markets were seen because of geo-political tension between US and North Korea. These observations need a closer examination from research perspective so as to examine what causes stock market to undergo extreme movement. This in turn can be used to design strategies for mitigating risks in the portfolio as well as cashing on opportunities.

Academically, as a researcher it gives an opportunity to understand the stock market movement to information that is not necessarily quantitative in nature. In addition, this also gives an opportunity to learn and examine the viability of news sentiment in understanding the movement of stock market.

1.4.2 Significance of the study

It has been a challenge for researchers to find out what kind of information creates movement in stock markets, by how much and how fast? Early studies on stock market movement employed quantitative data mostly in the form of dividends, earnings, economic indicators, etc. In recent years Gidofalvi (2001), Ferguson et al. (2011), Mo et al. (2016), etc. showed evidence for influence of news sentiment obtained from company-specific news articles on their related stock's variables like returns, volatility, trading volume, etc. It is also observed that stock markets tend to show abnormal movement when there is a feeling of extreme pessimism or optimism due to domestic or global events. Extreme pessimism has been found to cause market crash while extreme optimism has shown extreme exuberance leading to sudden rise in market prices. While there are instances when the feeling of pessimism or optimism continued for days driving market in the direction of its impact, there are also instances of markets recovering on the next day indicating a short term impact of pessimism or optimism. Identifying such short term extremes in market is important as investors or traders may get caught in a situation where they might fear the market to go down further and might sell their stocks by booking losses. On the other hand with market reaching higher levels due to positive sentiment, investors or traders may think that market may go up even further and buy stocks only to find the market correcting on the next day. So, it is important to identify such extreme market movement caused by sentiment. The sentiment itself is a result of events happening in the global and domestic market environment. Some of these events create a lot of buzz in the electronic media thereby generating sentiment having potential to affect investor's or trader's decision making. So far there is not enough research conducted using sentiment analysis on the content published in the news media to examine their influence on the stock market. While there are some events which repeat, a large number of events are unique and hence do not repeat, which makes it difficult for researchers to statistically prove the cause and effect relationship. One parsimonious approach to simplify this problem is to classify news about these events into categories, derive aggregate as well as category-wise news sentiment from them and then examine them for their influence on the stock market. This approach has been followed in this thesis focusing on the influence of news from business, economy, politics, and international categories during extreme movement in stock market.

This research is significant to academic fraternity and stock market practitioners with regards to the following contributions of the study:

1. In this research, theoretical foundations of Efficient market hypothesis (EMH) and noise trader theory are brought into the limelight providing an evidence on how these two theories complement one another. The evidence provided in this research suggests that stock market movements are caused according to Efficient Market Hypothesis but are subjected to turbulence because of noise traders' participation who are influenced by news sentiment thereby stretching prices to extreme levels for a short period of time.

2. This research has been built on the foundations of noise trader theory that considers limits to arbitrage and sentiment as the main contributors in moving stock markets to extreme levels. While the term "sentiment" used in this theory particularly refers to the behavioral biases of noise traders, this research has thrown light on news sentiment as a factor carried by news having influence on the emotions of noise traders.

3. Aggregate news sentiment derived from variety of news was not studied in the past literature. This research introduces the concept of aggregate news sentiment and discusses significance of it in the context of noise traders' reaction to sentiment aggregated from variety of news emanating from different domestic and global events.

4. This research extends the literature on news sentiment by incorporating news sentiment obtained from news categories like business, politics, economy, and politics; the extant literature had focused only on the business news category specific to company news sentiment and its influence on their related stocks. Thus, this research is a generalized form of news sentiment having influence on a portfolio of stocks rather than specific stocks.

5. While the extant literature has studied news sentiment focusing on individual stocks when news related to those stocks is published, there is not enough evidence explaining how

portfolio of stocks behave in response to news having significant impact on Country's macro-environment. This study has attempted to understand how news from business, economy, politics and international categories influences various sectoral indices when there is extreme movement in stock market.

6. This research is useful for different stakeholders operating in the stock market which include traders and investors, stock brokers, regulators, etc. If live aggregate news sentiment and category-wise news sentiment is made available on the trading platform which investors and traders use, it will help them to make strategy based on the news sentiment. News sentiment of each news category or aggregate news sentiment can be further used by brokers along with fundamentals of the stock or the technical indicators to provide recommendations to their clients.

7. This research also provides fund managers and retail investors an indication that when stock market deviates to extreme levels, it is not advisable to go for sector rotation strategy as every sector is influenced by aggregate news sentiment.

In short, significance of this study is to reinforce applicability and the role of noise trader theory to explain the extreme deviations in the short run influenced by aggregate news sentiments.

1.5 Scope of the Study

The scope of the study is confined with respect to the following:

- Data considered in this study spans over a period of 12 years starting from 1^{st} January 2008 to 31^{st} December 2019.

- This study uses secondary data in the form of daily adjusted closing prices of Nifty50 index and 10 Sectoral indices viz. Nifty Auto, Nifty Bank, Nifty Energy, Nifty FMCG, Nifty IT, Nifty Media, Nifty Pharma, Nifty Private Bank, Nifty PSU Bank and Nifty Realty.

- This study focuses on news sentiment obtained from 42,485 archived news articles from online news portal "economictimes.indiatimes.com".

- Only specific news categories are considered in this study which include Business, Economy, International and Politics.

- This study examines short term influence of news sentiment within an observation window of 5 trading days.

1.6 Overview of Methodology

In this thesis, aggregate news sentiment obtained from news belonging to politics, economy, international and business categories is examined for its influence on returns of stock market during extreme stock market movement. For obtaining trading days with extreme market movement, adjusted closing prices of Nifty 50 index in the period from 1st January 2008 to 31st December 2019 are chosen. Archived news articles from online news portal "economictimes.indiatimes.com" is used to webscrape articles during 5-day observation window[2] around the days of extreme stock market movement. This data is then used further to conduct analysis. Steps used in the methodology are briefly explained below:

1. Identification of trading dates with extreme returns: Since the study focuses on data points corresponding to extreme stock market movement, Nifty50 index returns are filtered during the above mentioned period to get trading days with extreme returns. This is done by setting a cut-off at +/- 2.58 standard deviation (sigma). Since volatility in the Nifty50 return series is not constant, standard deviation is estimated on a daily basis using GARCH(1,1) model and is used to set the cut-off at +/- 2.58 sigma. Also returns of other trading days in the observation window are considered along with extreme returns for further analysis.

2. Webscraping and News categorization: In order to derive the independent variables required in the study, news articles are webs scraped from the archives of "economictimes.indiatimes.com" news portal on all days in the observation window. They are then filtered and classified into business, economy, politics and international categories. News articles are also web scraped on non-trading days in the observation window.

3. Sentiment extraction and aggregation of news sentiment: All news articles webscraped are in html format, so they need to be converted to plain text, cleaned and pre-processed and stored as a corpus before using them for sentiment analysis. Lexicon method using Henry's Finance dictionary is applied to derived news sentiment polarity of each article. Aggregation of news sentiment is done by taking average of each news article's sentiment polarity over daily as well as across news categories.

4. Examination of influence of aggregate and category-wise news sentiment on returns of Nifty50 index: The influence of aggregate news sentiment on returns of Nifty index returns is examined using linear regression. Also, influence of news sentiment from business, economy, politics and international news categories on returns of Nifty50 index is examined using multiple regression. Spillover effect of news sentiment of previous two days on returns of day 3 is also examined.

[2] 5-day observation window includes the day of extreme stock market ovement, two days before and two days after the day of extreme stock market movement

5. Examination of influence of aggregate and category-wise news sentiment on returns of select sectoral indices: The influence of aggregate news sentiment on returns of ten sectoral indices viz. NIFTY Auto, NIFTY Bank, NIFTY FMCG, NIFTY Energy, NIFTY IT, NIFTY Media, NIFTY Pharma, NIFTY Private Bank, NIFTY PSU Bank and NIFTY Realty is examined using linear regression during extreme movement in Nifty50 index. Also, influence of news sentiment from business, economy, politics and international news categories on returns of these sectoral indices is examined using multiple regression during extreme movement in Nifty50 index.

1.7 Organisation of chapters

This thesis consists of 5 chapters organized as follows:

Chapter 1 provides the background of the topic followed by statement of the problem, objectives, motivation, significance and scope of the study. The chapter concludes by presenting an overview of the methodology and organization of chapters that provides a roadmap of this thesis.

Chapter 2 reviews the literature to provide the reader perspective on the theoretical foundations of how information and stock prices are related and also provides recent progress that has happened in the research related to news sentiment. To begin with, this chapter reviews literature on the three fundamental theories related to classical finance viz. Dow Theory, Random Walk Theory and Efficient Market Hypothesis. This is followed by literature review related to behavioral finance that includes understanding the meaning of investor sentiment, investor sentiment models based on psychological and sociological biases of investors, and empirical evidences of investor sentiment and its influence on stock market. This chapter then reviews literature on news sentiment to provide the reader an idea about news sentiment, various approaches researchers follow to conduct sentiment analysis, empirical findings of various studies related to news sentiment and sectoral indices. The chapter concludes by finding gaps in the literature followed by research questions, objectives, operational definitions and hypotheses of the study.

Chapter 3 presents the research methodology used in this thesis. It includes details about data collection method used to gather data, description of the econometric, statistical, and text mining techniques used to answer research questions. This chapter concludes by presenting the method used to process and analyze the data used in this thesis.

Chapter 4 presents the analyses and findings of the research conducted in this thesis and is divided into 5 sections that examines relationship between aggregate news sentiment on stock market returns, spillover effect of aggregate news sentiment on stock market returns, relationship between news sentiment of news categories on stock market returns, relationship between aggregate news sentiment and returns of sectoral indices, and relationship between news sentiment of news categories and returns of sectoral indices respectively.

Chapter 5 reports conclusions derived from the analyses conducted in this thesis including theoretical contributions, managerial implications, limitations of the study and pointers to related future research areas.

Chapter 2

Literature Review

2.1 Introduction

Stock market and its dynamics has always been an interesting area for research across the globe. It is a market where company stocks are listed and traded on stock exchanges. It mobilizes funds for companies through the issue of their equity shares in the primary market. Once listed, these shares (stocks) are then traded in the secondary market through stock exchanges. Investors buy and sell them so that they get returns. One of the ways to get returns, an investor has to see that they buy these stocks when they are underpriced and sell them when they are overpriced. However, knowing when stocks are underpriced and when they are overpriced is something investors have always tried to seek answer for. In this quest, researchers have come up with various theories and models, some of which argue that prices are not predictable while others have provided empirical evidence in favour of predictability of prices to some extent.

If we consider predictability of stock price movement, the past research has provided market professional with two broad approaches which they use for taking positions in the stock market and are popularly known as Technical analysis and Fundamental Analysis.

Fundamental Analysis: Fundamental analysis is used to calculate the true intrinsic value of a stock and then take positions in the market by looking at the difference in stock's intrinsic value and its current market price (Petrusheva and Jordanoski (2016)). Isidore et al. (2018) found that intrinsic value of a stock depends on various factors related to company, industry and economy; some of these important ones are given below:

- Company analysis: In case of company analysis, analysis is done on fundamental information about company's Earnings per share (EPS), Price/Earnings ratio (P/E), Return on Equity (ROE), Return on Assets (ROA), Debt/Equity ratio, Market capitalization, Price/Book ratio , Price/Sales ratio (Isidore et al. (2018))

- Industry analysis: With regards to industry analysis, factors like competition level, foreign entrants, government attitude, threat of potential entrants, cost structure, etc. are considered.
- Economy analysis: Economy analysis includes economic indicators like inflation, interest rates, purchasing power, growth rate, GDP, etc.

Technical Analysis: Technical analysis is based on Dow Theory and assumes that stock price movement can be predicted using the trend shown by past values in the time series. It does not care about the intrinsic value of the stock, but instead focuses on the price patterns created by supply and demand as well as the trading volume. While fundamental analysis is mostly used for long term investment decisions, technical analysis is preferred for short and medium term trading or investment decisions. One of the most popular charting tools to identify patterns is by using candlestick chart which displays high, low, opening and closing prices on a candlestick at a given trading interval. Some popular patterns which are used to take buying and selling decisions include “head and shoulder" pattern, “double top" pattern, “double bottom" pattern, etc. Apart from candlestick patterns, technical analysis has a wide variety of indicators like moving averages, Relative Strength Index (RSI), MACD (Moving Average Convergence Divergence), Bollinger bands, etc. which provide direction as well as entry and exit levels for taking positions in stocks.

While fundamental analysis and technical analysis are the two most popular approaches among the market professionals, the research community has explored stock market from different perspectives. Right from developing asset pricing models, examining market efficiency, understanding investor behaviour to the more recent work using event study methodology, sentiment analysis, etc., stock markets have given researchers opportunities to theorize and provide empirical evidence of their work. This in turn has benefited market participants, practitioners, regulators and the change agents in the stock markets. Stock markets have evolved over a period of time and so has been the research conducted on it. While it was classical finance that provided initial models, theories, and empirical evidences, behavioural finance complemented them in the recent years to make them more applicable in the real market setup.

This chapter reviews the past literature to familiarize ourselves with the research that has been conducted to investigate the influence of information on stock markets. To begin with, the literature review discusses the classical finance theories that form the basis for predicting stock market movement using information. This is followed by the foundation theories of behavioral finance that attempted to explain behavior of stock market movement which the classical finance theories found difficult to explain. These behavioral finance theories lead us to the studies that focused on “Sentiment" and its role in stock market movement. The focus of this research being news sentiment and its influence on stock market, past studies on news sentiment, sentiment analysis techniques and other related studies have been discussed in detail in this chapter.

The chapter concludes by identifying gaps in the extant literature based on which further research is conducted.

2.2 Classical finance approach to stock price movement

Classical finance research covers a wide spectrum of studies on stock market which includes theories and models for determining asset prices, portfolio selection, option pricing, movement in asset prices, etc. It can be broadly classified into two themes; valuation of asset and asset price movement. Among various studies, the two most popular models that deal with the valuation of assets are the Capital Asset Pricing Model and the Arbitrage Pricing model. On the other hand, research on stock price behavior has evolved over a period of time. Significant contribution to this strand of research has come from Random walk theory, Dow theory, Efficient Market Hypothesis and Noise trader theory. Both these themes of research have formed the basis for practitioners' strategies in the real market setup. Since, in this thesis, the focus is on the behavior of stocks rather than the valuation of stocks, the three most widely applied theories from classical finance viz. Dow theory, Random walk theory and Efficient market hypothesis are discussed below:

2.2.1 Dow theory

Dow theory originated from the ideas of Wall Street Journal's first editor Charles Dow. He invented two stock market averages; Dow Jones Industrial Average (DJIA) and Dow Jones Transportation Average (DJTA) and are part of his work on understanding stock market movement. The Dow Jones Industrial Average also known as the "Dow 30", is a stock market index created back in 1896 by Dow Jones and his business partner Edward Jones that tracks 30 large, publicly-owned blue chip companies trading on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ). Dow Jones Transportation average also known as "Dow Transports" is a price-weighted average of 20 transportation stocks traded in the United States and is the oldest U.S. stock index, first compiled in 1884 by Charles Dow.

Dow's ideas were aggregated by his successor William Hamilton and other followers like Robert Rhea, Nelson , E. George Schaefe, etc. and was called the Dow theory. The six principles of Dow theory as cited by Edwards et al. (2018) are given below:

1. The averages discount everything: Since Industrial average and Transportation average covers stocks representing market activities of large number of investors including the ones having best of the information on trends and events, these averages reflect all information about the events that has happened and the ones likely to have influence on both demand and supply.

2. The market has three trends: These are primary, secondary and minor trend. The primary trend is the longest, lasting for more than a year and could be either rising (bullish) or falling

(bearish). Secondary trend is an intermediate trend that represents corrections in primary trend and lasts for three weeks to three months retracing from one-third to two-thirds of the previous trend movement. Minor trend is a short-term movement usually lasting for less than three weeks and is related to the secondary trend.

3. Major trends have three phases: Every primary (major) trend consists of three phases viz. accumulation phase, public participation phase and distribution phase. In case of bullish period, accumulation phase represents reviving confidence in the future of business, public participation represents the response of stock prices to the known improvement in corporate earnings and distribution phase represents a period speculative in nature with stock prices getting advanced on hopes and expectations. In case of bearish period accumulation phase represents abandonment of hope resulting from buying at inflated prices, public response represents selling because of decreased business and earnings and distribution phase represents distress selling of sound securities irrespective of their value.

4. The averages must confirm each other: For confirming a bull market, both the averages viz. transportation average and industrial average must have begun a bull trend. The precise point for confirmation is when the second average moves above a previous high after making a higher low. Similarly, for confirming a bear market, both averages must have begun a bear trend. The precise point for confirmation is when the second average moves below a previous low after making a lower high.

5. Volume must confirm the trend: An overbought market becomes dull on rallies and active on decline. Similarly an oversold market becomes dull on declines and active on rallies.

6. A trend is assumed to be continuous until an opposite trend commences its reversal: Since very precise conditions are required for the ending of a trend and the start of a new one, it should always be assumed that the current trend is still in effect until the opposite trend is definitely in place.

In summary, Dow theory explains the stock market movement using trends shown by transportation average and industrial average. Since these averages are obtained using past information (prices), it indirectly indicates the use of past information in forecasting the future prices. Compared to its counterparts, this theory explains the movement in stock market using transportation and industrial average based on the supply and demand for different stocks which are constituents of these averages. This theory thus forms the basis for technical analysis used by practitioners who use the trends in past prices to predict future movement and doesn't give much attention towards the information related to the fundamental factors affecting the real value of the stock. During three phases of major trend which happen during bull and bear period, investors go through emotions of hope and distress.

2.2.2 Random Walk Theory

Random walk theory originated from the work of Louis Bachelier explains the movement of stocks using statistical trend called as “random walk”. Random walk theory is a diametrically opposite to technical theories which assume that historical patterns in the series is a good estimate of their future behaviour. According to random walk theory, changes in stock prices have the same distribution and are independent of each other. According to Davis (2006), the earliest form of random walk theory was presented by Louis Bachelier in 1900 in his PhD thesis “Theorie de la Speculation" in relation to bond market prices. However, economic side of Bacheliers work was completely ignored for a long period of time until it was taken up by Paul Samuelson in the 1960s.

Fama (1965) posited that an efficient market at any given point in time reflects information about the past events as well as the events expected in the future into prices. However, disagreement among various market participants about what the intrinsic price of the security makes the actual price to wander randomly about its intrinsic value. This randomness in series of stock price changes has no memory i.e. they are independent and hence past history of the series cannot be used to predict the future.

Malkiel (1973) further propounded that stock prices follow random walk pattern and hence short-run changes in stock prices cannot be predicted, making investment advisory services, earnings predictions, and complicated chart patterns useless.

Random walk theory, in short laid foundations for the strand of research that gave no importance for past information with regards to its use in predicting future values.

2.2.3 Efficient Market Hypothesis

Efficient Market Hypothesis (EMH) proposed by Fama (1965) has been one of the most popular, path-breaking and yet highly debated theories researched globally across various stock markets. It simply suggests that stock prices fully reflect all available information (Fama, 1970) and that it is difficult to generate economic profits from a stock by using all information available about it (Jensen, 1978). According to Timmermann and Granger (2004), logic behind EMH is that if returns were predictable, investors would generate unlimited wealth which for a stable economy was impossible. Fama (1965) defined efficient market as: “a market where there are large numbers of rational profit maximizers actively competing, with each trying to predict future market values of individual securities, and where important current information is almost freely available to all participants"

The sufficient conditions for a market to be considered as efficient include:

- There are no transaction costs for trading securities.
- All market participants have equal access to historical data on stock prices as well as public and private information.

- All agree on the implications of the current information for current price and distributions of future prices of each security.

Going to the roots of EMH, early foundation for EMH was laid out when empirical evidences were provided for random walk theory. Hypothesis of random walk theory that future steps cannot be decided based on past steps has a close resemblance to EMH that states prices fully reflect all available information.

Categories of Market Efficiency: Fama (1970) classified market efficiency into weak, semi-strong and strong depending on the kind of information reflected in the prices. If prices reflected information about historical prices, market is considered to be having weak form of efficiency. If all publicly available information like announcements related to earnings, stock split, new security issues, etc. is reflected in the prices, market is semi-strong efficient whereas if security prices fully reflected privately held or information having monopolistic access then the market is considered to be having strong form of efficiency.

EMH - a challenge to fundamental analysis and technical analysis: EMH had posed a serious challenge to those who felt forecasting can be done using fundamental analysis as well as technical analysis. EMH in its semi-strong form made the fundamental analysis vulnerable because of its dependency on information related to performance of a company in terms of earnings, dividend, book value, etc. In the same manner, EMH in its weak form makes technical analysis vulnerable because of its dependence on historical prices to forecast future prices. This tussle between researchers from classical finance and the market professionals was put into perspective by Gilson and Kraakman (2003) giving a general explanation for the elements that lead to and limit market efficiency in their thesis known as "Mechanism of Market Efficiency" (MOME). They also coined the term “relative efficiency” to mean that particular information might be reflected in real as opposed to ideal market prices more or less rapidly. If it gets rapidly reflected in prices, it represents theorist’s model of market populated exclusively by fully-informed traders. Thus, relative efficiency looks at how rapidly information get subsumed into prices rather than how correctly the market prices are close to the expected prices obtained using asset pricing model like CAPM. Gilson and Kraakman (2003) further classifies information-based trading activity into four categories taking into account ability of the market to incorporate information into market prices with progressively decreasing relative efficiency.

(a) Universally-informed trading: Market prices immediately reflect information as all traders know about it; representing theorists’ perfect market assumption.

(b) Professionally-informed trading: Information widely not known to public that is rapidly reflected in prices through the trading of savvy professionals.

(c) Derivatively-informed trading: Inside information known to a very few traders will be reflected slowly into the prices as uninformed traders will not be aware of this information but will be active by observing the activities of presumptively informed traders or unusual price and volume movements

(d) Uninformed trading: Information known to no one might be reflected slowly and imperfectly into the price due to aggregation of forecasts of numerous market participants with heterogeneous information

Critics of EMH: For almost two decades from 1960s to 1980s, EMH was considered to be one of the pillars of stock market research. Jensen (1978) even stated that there was no other proposition in economics which has more solid empirical evidence supporting it than the EMH. However, in the early 1980s and thereafter, it faced some serious criticisms particularly from the proponents of behavioral finance. Prechter and Parker (2007) highlight several shortcomings in EMH like failure to explain financial market valuation, failure to consider relevant aspects of human behavior, mean reversion problems, excessive volatility problem, calendar-based anomalies such as the January effect, joint hypothesis problem and dependency on the theories like random walk model and martingale model which are also criticized for their unrealistic assumptions.

According to Fama (1991), market efficiency alone is not testable, it has to be jointly tested with some model of equilibrium pricing (asset pricing model) and many such models exists. If one finds an anomalous evidence on behaviour of returns, this ambiguity is split between equilibrium model used as reference against which abnormal returns are tested and the market efficiency model. That eventually makes EMH untestable.

Baker and Wurgler (2007) claimed that standard finance models which focus on rational behaviour of investors cannot explain extreme movement in stock markets. For instance, they fail to explain extreme market movement like Nifty Fifty bubble of the early 1970s, Black Monday crash of October 1987, the Internet or Dot.com bubble of the 1990s, etc. Following De Long et al. (2003) and Barberis et al. (1998), they suggested that the standard finance model should be augmented by making the two basic assumptions: (1) Investors are subject to sentiment and (2) There are limits to arbitrage; betting against sentimental investors is costly and risky because of which rational investors and arbitrageurs find it difficult to get the prices back to their fundamental level. They find speculative stocks to be more sensitive to investor's sentiment due to difficulty and subjectivity of determining their true values.

Referring to the “bounded rationality" model proposed by Herbert Simon, Barber and Odean (2008), Hong and Stein (1997) criticized EMH for not being realistic with the assumptions. They argued that standard economic and financial models are not complete and accurate as they assume perfect information accessibility and processing capabilities, decision making alternatives and consequences of actions. In reality, humans have limited processing capabilities which makes it difficult for them to select a particular decision choice from a large set. However, in support of EMH, Fama (1991) stated that even though because of the joint-hypothesis problem precise inferences about the degree of market efficiency are likely to remain impossible, EMH needs to be judged on how it has improved the understanding about the behavior of security returns. In this review, he pointed out that two decades after the introduction of EMH, weak form of efficiency broadened a general coverage on tests for return predictability using variables like dividend yields and interest rates rather than just testing forecasting power of past returns. Similarly, semi-strong form and strong form had translated into event studies[1] and test of insider information respectively. Short horizon and long horizon event studies along with cross sectional studies also emerged based on the foundations laid by EMH.

Literature on EMH reviewed above reveals that stock market movement as described in the original form of EMH (Fama (1970)) was too rigid because of its unrealistic assumptions. However, issues raised by practitioners and the behavioral finance researchers about EMH improved the knowledge on stock market research making it more practical. This paved the way for a large number of event and sentiment related studies on various stock markets globally.

2.3 Behavioral finance approach to stock price movement

Many popular theories in Classical finance including EMH is built on the premise that rational investors have access to all information and they evaluate all possible outcomes while taking decision. Behavioral finance researchers like Barber and Odean (2008), Barberis et al. (1998), Daniel et al. (1998), Fenzl and Pelzmann (2012), Hong and Stein (1997), etc. criticized EMH by counter argument that these assumptions do not hold good in practice. They further pointed out that such classical finance theories have no answer for extreme market movements and do not provide any empirical evidence for market regularities like “underreaction”, “ overreaction” and “momentum”. On the contrary, they find that many of the explanations for such phenomena lie in studies related to psychology of human behaviour and their decision making. For instance, the psychological and sociological patterns of human behaviour studied by Kahneman and Tversky (1979) and psychologist Edwards in 1968 were found useful in behavioural finance research to explain these regularities.

The arguments presented to criticize EMH with supporting evidences for anomalies and other patterns of stock market movement pointed towards importance of behavioral characteristics of investors. Studies on investor sentiment came into the limelight and many researchers including Baker and Wurgler (2007), Barberis et al. (1998), De Long et al. (2003), Mian and Sankaraguruswamy (2012) contributed in explaining how investor sentiment influences stock market movement. In order to understand the role of investor sentiment one has to contrast the

[1]An event study measures the impact of a specific event on the value of a firm and has been applied to a variety of firm specific and economy wide events (MacKinlay, 1997)

classical finance perspective on investor sentiment vis-a-vis behavioral finance perspective. According to C. Zhang (2008), classical finance provides no room for investor sentiment since price changes only reflect the arrival of external news about future cash-flows and interest rates where as the behavioral finance suggests that investor sentiment may also significantly distort market outcomes thereby affecting asset prices in equilibrium.

Even though behavioral finance theories and models came into existence as a result of the criticisms faced by classical finance theories like EMH, they were definitely not considered as replacement of classical finance theories, but rather they augmented them to provide better explanation for stock market movement. Research in behavioral finance was able to answer questions which classical finance theories found difficult to answer. For example, questions like how individual investors pick up their portfolio compared to institutional investors? how do individual investors react based on their expectations about future earnings? how herd behaviour of investors impact stock market? These questions found prominence in behavioural finance research like Barber and Odean (2008), Bikhchandani and Sharma (2011), Hong and Stein (1997), etc. that used psychological and sociological bias of investors on stock market.

With investor's role in stock price movement getting prominence among the then researchers globally, behavioral finance research had gained a momentum and many researchers wanted to explore it even further. Some of the themes that were popular among those researchers are discussed below:

1. Investor sentiment: definition and measurement
2. Theoretical foundations of behavioral Finance
3. Models of investor sentiment using psychological biases of investors
4. Models of investor sentiment using social biases of investors

2.3.1 Investor sentiment: Definition and measurement

Definition of investor sentiment

Literature on behavioral finance uses the term "investor sentiment" quite often but do not seem to agree on any universal definition. Referring to the market participants' sentiment, Kirkpatrick and Dahlquist (2011) defined it as the net amount of any group of market players' optimism or pessimism reflected in any asset or market price at a particular time. Baker and Wurgler (2007) defined investor sentiment as a belief about future cash flows and investment risks that is not justified by the facts at hand. This definition reflects subjectivity in the interpretation of the information received by an investors based on the beliefs they holds. More specifically, according to C. Zhang (2008), investor sentiment represents market participants' potential erroneous beliefs about future cash flows relative to the true fundamental value of the underlying asset.

Fundamental value here refers to the discounted sum of future cash flows and investment risks. The definition of investor sentiment given by Baker and Wurgler (2007), C. Zhang (2008) suggests two possibilities of how investors use information. One possibility is that the investors correctly use wrong information (For e.g., see C. Zhang, 2008) and the second possibility is that they wrongly use the correct information (For e.g., see Barberis et al., 1998). According to C. Zhang (2008), investor sentiment is the difference between two forecasts made by the investors viz. the subjective assessment using all available information in addition to potentially biased private information and objectively correct and relevant information. Mathematically this can be expressed as:

$$S_{i,t} = E_{i,t}[P_{t+1}|I'_{i,t}] - E_t[P_{t+1}|I_t] \tag{2.1}$$

where $S_{i,t}$ denotes individual sentiment of investor i. P_{t+1} represents the stock price at time (t+1) whose true value is unknown at time t. I_t represents all public information available about fundamentals at time t, and $I'_{i,t}$ denotes the information actually used by individual investors to arrive at their forecasts.

Measurement of investor sentiment

Early studies in behavioral finance mainly focused on explaining why investor sentiment is one of the important variables in the stock market movement. However, the real challenge researchers faced was to measure it and then examine its impact on stock market. According to Baker and Wurgler (2007) there are two approaches for understanding how investor sentiment affects stock prices. Bottom-up approach uses biases in individual investor psychology, such as overconfidence, representativeness, and conservatism, to explain how individual investors underreact or overreact to past returns or fundamentals. To capture these, survey method is used. Top-down approach on the other hand measures investor sentiment using various sentiment proxies. Baker and Wurgler (2007) uses six standardized sentiment proxies after eliminating effect of macroeconomic conditions. These include the closed-end fund discount (CEFD), detrended log turnover (TURN), number of IPOs (NIPO), first-day return on IPOs (RIPO), dividend premium (PDND), and equity share in new issues (S). They create two measures of investor sentiment- sentiment level index $SENT$ and sentiment change index $\Delta SENT$ to test for return predictability conditional on the state of sentiment and return co-movement patterns associated with changes in sentiment respectively. These two sentiment measures are given by the following equations:

$$SENT = -0.23CEFD + 0.23TURN + 0.24NIPO + 0.29RIPO - 0.32PDND + 0.23S \tag{2.2}$$

$$\Delta SENT = -0.17\Delta CEFD + 0.32\Delta TURN + 0.17\Delta NIPO + 0.41\Delta RIPO - 0.49\Delta PDND - 0.28\Delta S \tag{2.3}$$

Bandopadhyaya and Jones (2006) developed Equity Market Sentiment Index (EMSI) as a representative of investor sentiment and demonstrated how this measure relates to stock price movements of a group of firms from Massachusetts Bloomberg Index (MBI). EMSI captures the shift in risk attitude of investors which is a driver for changing stock prices. EMSI is computed as follows:

$$EMSI = \frac{\sum(R_{ir} - \bar{R_r})\sum(R_{iv} - \bar{R_v})}{[\sum(R_{ir} - \bar{R_r})^2 \sum(R_{iv} - \bar{R_v})^2]^{1/2}} * 100 \tag{2.4}$$

where:

R_{ir} is the rank of the daily return for stock i

R_r is the population mean return

R_{iv} is the rank of the historical volatility for security i

R_v is the historical volatility rankings

2.3.2 Theoretical foundations of behavioral finance

The core aspect of most of the behavioral finance theories lies in how investor sentiment explains the behavior of stock market for those issues which classical finance struggled to explain. For instance, Baker and Wurgler (2007) cite an example of 1990's internet bubble and a subsequent crash that happened because of investors' speculative tendencies in buying difficult-to-value technology stocks. There are several such examples of irrational exuberance in the history of global stock markets as well as several market anomalies that the classical finance theories found difficult to explain, but by augmenting them with research from behavioral finance this was possible.

Many of the classical finance theories including Modern Portfolio theory and Capital Asset Pricing theory which had gained popularity in the period around 1960s were based on the common premise that investors are rational. Rationality here implies that all investors are identical with a goal of utility maximization, they follow “Bayes rule" to form new beliefs as new information becomes available and possess ability to predict accurately. In other words, this means that the behaviour of investors do not have much role to play in deciding asset prices. Contrary to this, Kahneman and Tversky (1979) provide several classes of choice problems in which preferences systematically violate the axioms of expected utility theory. They claim expected utility maximization model is not an adequate descriptive model for decision making under risk in all situations and suggests Prospect theory where in people make decisions based on the potential value of gains and losses rather than the utility of the decision. There were many other behavioural finance researchers like Thaler and De Bondt (1985), Barberis et al. (1998), Bikhchandani and Sharma (2011), Baker and Wurgler (2007), etc. who criticized the rationality assumptions of classical finance theories because of their inability to explain mispricing in stock market, provided better explanation by relaxing the restrictive assumptions of classical finance models like EMH. Others like MacKinlay (1997), Quaye et al. (2016), Liew and Rowland (2016) and Geetha et al. (2011) explained influence of company specific event (earnings announcement), global event (Brexit), political event (election), macroeconomic event (inflation rate, exchange rate, interest rate and GDP) on stock market returns respectively. Piccoli et al. (2017) investigated variety of global and domestic events and found that stock markets tend to overreact after both positive and negative events, as well as global and domestic shocks causing extreme returns in stock market.

Noise Trader Theory

One of the most important behavioral finance theories that provide explanation for extreme deviation of prices is the Noise Trader Theory advanced by Black (1986), De Long et al. (2003) and Shleifer and Summers (1990). This theory classifies investors into two categories viz. the informed investors called arbitrageurs and the uninformed investors called noise traders. Informed investors trade using information, but they cannot predict when overvalued or undervalued stock prices will return to their fundamental prices. Noise traders on the other hand are influenced by the noisy signals or sentiments generated because of their irrational expectations about future demand and supply conditions within and across sectors Black (1986). When arbitrageurs trade against these noise traders they bear an additional risk called as "Noise traders' risk" apart from the fundamental risk. According to De Long et al. (2003), noise traders' risk is the risk arbitrageurs have to bear in addition to fundamental risk because of the noise traders' tendency to follow their beliefs that do not revert to their mean for a long time with a possibility of further becoming even more extreme. Because of such behavior of noise traders, arbitrageurs find it difficult to bet against them and limits the size of positions they take in the market. This results in asset prices significantly deviating from fundamental values even in the absence of fundamental risk and sometimes reach extreme levels. So, in short, according to this theory there are two factors viz. sentiment of noise traders and arbitrageurs' limits to arbitrage important that influences stock market to achieve extreme levels. While this theory explained the mechanism of large deviation in prices, the basis for why noise traders are prone to take decisions based on their sentiments can be traced in the models of underreaction and overreaction proposed by Barber and Odean (2008), Bikhchandani and Sharma (2011), Fenzl and Pelzmann (2012), Tversky et al. (1974), etc. which identify psychological and sociological biases in investors like conservatism, representativeness bias, etc. as important contributors to market phenomenon like underreaction, overreaction, momentum, etc.

A simple example to convey this idea would be to imagine a situation where market is continuously going up and very bullish. Looking at the way market has progressed over the past few days, investors who have been inactive for a while now buy stocks moving the price even further. With continuous rise in price, investors become overconfident and their greed tempts them to buy stocks irrespective of high value. This drives the market even further to a peak level where there is no money left with the investors to buy any more stocks. Because of this, upward movement of the stock price gets a halt. At this point, the investors who had bought the stock at the beginning of this rally and sitting pretty with handsome returns, they realize the opportunity for booking profit. They start selling stocks leading to a drop in stock price. As the prices start dropping, other investors who still have got a reasonable profit but don't want to lose it also start selling. With more and more selling happening, now they become pessimistic and fearful. Even the ones who are in a slight loss also start selling stocks in order to avoid further losses. As this happens, panic sets in among other investors too, including the ones who had bought these stocks at higher prices. All of a sudden, most of the investors start selling stocks in panic and the stock price crashes down quickly. So, here we see how investors' optimism and pessimism affects market movement and how in turn market movement creates these feelings in the investors.

Black (1986) provided explanation for noise and its role in noise trading. According to him, noise arises from a large number of small events, which is often a causal factor much more powerful than a small number of large events can be. The information released during such events is difficult to decipher from stock market's perspective and investors and traders have differences of opinion regarding such information. Quite often they misjudge this information and trade on noise thinking that it is information. This results in two types of traders: informed traders who trade on the information and noise traders who trade on noise causing inefficiency in the market. The presence of both types of traders in the market causes the stock prices to reflect both the information that informed traders trade on and the noise that noise traders trade on. While trades which happen based on information reflect prices with market efficiency, the trades which happen based on noise reflect prices with inefficiency causing deviations away from the prices reflected by trades based on information.

Compared to the printed media which was the main source of providing content to public a few decades back, online media today covers a large number of events and is the widely used media by public. With the massive volume of online content received by traders about the domestic and global events, it makes it difficult for them to filter information from noise. So, there is a high tendency among the traders to misjudge this content which increases the possibility of them using it as noise thinking that it is information.

Along with Noise trader theory, there were several models of investor sentiment proposed by researchers for underreaction and overreaction based on the psychological and sociological biases of investors; some of which are discussed below:

2.3.3 Models of investor sentiment using psychological biases of investors

According to Fenzl and Pelzmann (2012), studies on investor sentiment using psychological bias predominantly use a model of bounded rationality in economic behavior as given by the psychologist Simon. In contrast to rational models used by the classical finance theories, the bounded rationality model allows for lack of information, missing alternatives, uncertainty about endogenous or exogenous events, and the impossibility of taking all possible outcomes and consequences into account. So, within a dynamic market set up, investors react with their inherent beliefs to new information updates about endogenous and exogenous events. Barberis et al. (1998), Daniel et al. (1998), Hong and Stein (1997), Thaler and De Bondt (1985), etc. have used psychological biases in investors for explaining the role of investor sentiment in stock market phenomena like underreaction and overreaction. These models of underreaction and overreaction are discussed below:

Model of underreaction and overreaction by Barberis et al. (1998)

The model of investor sentiment proposed by Barberis et al. (1998) assume that investors form their beliefs from the information they received in the recent past and use them to take their decisions when they receive new information. Also there is a possibility of investors using the correct information wrongly but update their beliefs based on the outcome of their decision. Moreover, it is assumed that the earnings of the asset follow a random walk which the investor does not know. According to him firm's earnings moves between two “states" viz. mean-reverting and trending. Also, the transition probabilities between these two states and the statistical properties of the earnings process in each one of them are fixed in the investor's mind. In each period, the investor updates his beliefs and decides about the state based on earnings information. When a positive surprise is followed by another positive surprise, the investor decides the state as trending whereas if a positive earnings surprise is followed by a negative earnings surprise or vice versa, he decides the state as mean-reverting. This model of decision making by investors results in stock market phenomena such as “underreaction" and “overreaction" as explained below:

According to the model of investor sentiment by Barberis et al. (1998), stock prices are considered to have underreacted if average return on the companys stock in the period following an announcement of good news is higher than the average return in the period following bad news. That is, stock has shown underreaction to good news (a mistake), corrected in the following period giving higher returns. Suppose that in each time period, the investor receives news about a particular company and at time t, the news he receives is represented as z_t. If we represent good news by G and bad news by B, then the stock prices are considered to have underreacted if the average return on Company's stock in the period following an announcement of good news is higher than the average return in the period following bad news. Underreaction can be represented in the mathematical form as follows:

$$E(r_{t+1}|z_t = G) > E(r_{t+1}|z_t = B) \tag{2.5}$$

This finding obtained from the model of underreaction goes well with the conservatism phenomenon from psychology described by Psychologist Edwards in 1968 who states that individuals are slow to change their beliefs in the face of new evidence. When applied to investors, conservatism here means that investors partially update their prior information on news announcements. On the other hand, stock prices are considered to have overreacted if the average return following not one but a series of announcements of good news is lower than average return following a series of bad news announcements. This can be attributed to the fact that an investor after getting a series of good news becomes over-optimistic and assumes that future announcements will also be good and hence overreacts sending stock price to higher levels. However, when subsequent news announcements contradicts his optimism, it leads to lower returns.

Using the study of Tversky et al. (1974) on “representativeness heuristics", Barberis et al. (1998) explained overreaction in stock market. According to this, people following representative heuristics or thumb rules to evaluate the probability of an uncertain event, or a sample, by the degree to which it is (i) similar in its essential properties to the parent population (ii) reflects the salient features of the process by which it is generated. When applied to investor behaviour, this phenomenon describe situation in which investors assume past history as representative of the future growth potential undermining the fact that history of high earnings growth is unlikely to repeat itself; they overvalue the company, and get disappointed in the future when forecasted earnings growth fails to materialize. This results into overreaction in the market. This model of overreaction using representative heuristics can be better understood from the investor’s reaction to good news he received in the past. When an investor finds that there is a series of good news in the stock, he starts reacting to this by buying stocks. As good news repeats in the subsequent periods, he becomes more and more confident and reaches a stage where he now expects good news to arrive in the future as well. He becomes overconfident (irrational) and continues to buy stocks using his representative bias without waiting for news to arrive. This takes the stock prices to unprecedented high levels. Consequently, when the news arrives and turns out to be a bad one, he then rationally sells stocks. When many investors show such behavior at the same time, this gives rise to “overreaction” which can be mathematically expressed as follows:

$$E(r_{t+1}|z_t = G, z_{t-1} = G, \ldots\ldots, z_{t-j} = G) < E(r_{t+1}|z_t = B, z_{t-1} = B, \ldots\ldots, z_{t-j} = B) \tag{2.6}$$

Thus during overreaction, the average return following not one but a series of announcements of good news is lower than the average return following a series of bad news announcements.

In today’s digital information age, investors have access to information via social media and news media. Particularly, many of the news articles are written in such a manner that provide a context of the past as well as what is happening today. This context as well as the tone in news article can be a fodder to their representative heuristics thereby inducing overreaction as same news is available to large section of investors.

Model of underreaction and overreaction by Hong and Stein (1997)

In the unified model of underreaction and overreaction Hong and Stein (1997) explained how underreaction leads to overreaction followed by reversal. In this model there are two types of agents “newswatchers” and “momentum traders”. The model assumes that: (i) The newswatchers make forecasts based on their private signals they observe about future fundamentals and they disregard current or past prices. (ii) Momentum traders use only changes in past prices to forecast. (iii) Private information diffuses gradually across the newswatcher population. Based on these assumptions it follows that initially (at t=0), when only newswatchers are active, prices adjust slowly to new information exhibiting underreaction but never overreaction. As momentum traders base their forecast on this gradual increase in price and enter at (t+1), they force the price to go higher resulting in some profit. However, later momentum buyers entered at time (t+i for some i) lose money as at that time price go far above the long-run equilibrium and reversal happens.

This model suggests that news which is public information acts as an initial trigger in terms of identifying the direction of movement but the actual strength in the movement comes from the participation by momentum traders based on the changes in price. Thus, news here becomes a source of influence creating an initial movement in price building expectations in the momentum traders who are influenced by the change in price. Since, there is an underreaction followed by overreaction, this phenomenon can be considered as a manifestation of semi-strong efficiency.

Model of overreaction by Daniel et al. (1998)

Model of overreaction by Daniel et al. (1998) is based on two psychological regularities: overconfidence and attribution bias. Based on the evidences of their study, they posits that investors overestimate their abilities and view themselves as having better abilities to value securities as compared to what others think about them and underestimating their forecast error variance. This leads to overconfidence. Secondly, this confidence grows when public information is in agreement with his information, but it does not fall commensurately when public information contradicts his private information thus exhibiting self-attribution i.e. people tend to credit themselves for past success, and blame external factors for failure. The effect of overconfidence is that overreaction drives the prices higher than the fundamental value and thereafter a long run reversal occurs.

In today’s context. with public information available online, overreaction as described in this model can happen in following ways: (i) Online news creating an environment that matches the beliefs of the investor. (ii) Investor finding his beliefs in sync with analysts recommendations published in online media.

Model of overreaction by Thaler and De Bondt (1985)

Thaler and De Bondt (1985) explained overreaction effects as follows: “If stock prices systematically overshoot, then their reversal should be predictable from past return data alone, with no use of any accounting data such as earnings”. They suggested two hypotheses: (a) Extreme movements in stock prices will be followed by subsequent price movements in the opposite direction (Directional Effect) (b) The more extreme the initial price movement, the greater will be the subsequent adjustment (Magnitude Effect). Their study gave empirical evidence of overreaction by creating winner and loser portfolios each consisting of 35 stocks from NYSE and then measuring the overreaction over long periods (3 or more years) by computing cumulative average returns. Loser portfolios of 35 stocks were found to outperform the market by, on average 19.6%, thirty-six months after portfolio formation. Winner portfolios, on the other hand, earned about 5.0% less than the market. So the difference in cumulative average residual between the extreme portfolios, equaled 24.6 % (t-statistic: 2.20).

Brown and Harlow (1988) included intensity effect in overreaction which is stated as “shorter the duration of the initial price change, the more extreme the subsequent response".

2.3.4 Models of investor sentiment using social biases of investors

Investors are not only subjected to their personal psychological bias but they may also get influenced by the social interactions which they have with others. This is very important with respect to how multiple investors behave in stock market. An individual having an ability to create a strong influence on others may create a tendency in them to make decisions in one particular direction. For instance, recommendations given by stock market analyst or a successful investor are mostly followed by many other investors. Behavioural finance researchers claim that such social bias may have a significant contribution to market bubbles and crashes. Mechanism by which individual investors get influenced by such social bias is explained below:

Model of investor sentiment by Fenzl and Pelzmann (2012)

Fenzl and Pelzmann (2012) suggested that in complex and uncertain situations, people tend to look at how other people behave and then realign themselves to evaluate the best course of action. Sometimes because of influence of collectives, people tend to follow each others’ mistakes by way of rumours or speculations and critical thinking and doubts are not given any attention. They define social contagion as the spreading or transmission of social or psychological influence in direct or indirect contacts between individuals and their environment. Sometimes, an individual may not be directly involved in the action but he may act like a carrier. For example, an investor A gets some news about a stock but he doesn’t participate in the market based on this news. However he conveys this news to investor B who participates in the market. Here, investor A acts like a carrier of information. In a contagion process some agents may have more influence on others. Also, this relationship of influence is asymmetric, may change in magnitude based on the number of exposures in a certain time frame and depends on susceptibility of contagious entity. Social contagion is most likely to happen when investors share their views about stock movement pertaining to some expectations from an event related to company or events having national or global impact. It is also possible when stock market analyst express their outlook on certain stocks or provide stock recommendations.

Model of investor sentiment by Bikhchandani and Sharma (2011)

According to Bikhchandani and Sharma (2011), a market participant can be said to have herd if he decides to do an investment without knowing other investors' decision but doesn't do so when he comes to know what others have done. They find three causes for "herding - (1) Others may know something about the return on the investment and their actions reveal this information. (2) Money managers imitate others because the incentives provided by the compensation scheme and terms of employment may be such that imitation is rewarded. (3) Individuals may have an intrinsic preference for conformity in order to avoid the blame of being different from others. Herding can also have a snowball effect wherein initially few individuals may be part of herding which later grows significantly because of such imitation. Similar to herding, in a sequential decision making setup, they also find a phenomenon called Information cascading that may occur when individuals override their private information and use what is believed to be informative public signals for decision making.

Herding and social contagion are particularly important in understanding extreme exuberance in stock market. It can be used to explain why investors think alike in their decision making in more or less same time frame and thereby making prices to move away from equilibrium levels. As herd, social contagion and information cascading phenomena happen as a result of interaction between various investors or non-investors, today's social media and news media plays the role of enabler. While news media provide latest updates about what is happening around the word, social media allows users to express their opinion on events. An investor on social media may be able to understand opinion of others on a particular topic or an event which may have direct or indirect relevance to stock market. If many of the participants share the same opinion, it is more likely to create a bias in a particular direction which may be seen by participants as the most correct opinion, even though their personal beliefs may be different. This bias towards other's actions or opinions may be likely to take place if someone has a tendency towards herding rather than believing in himself.

2.3.5 Empirical evidence on the reaction of stock market due to investor sentiment

Research in classical finance was able to provide an explanation for an ideal behavior of stock market. However, it was not able to convincing explain certain regularities in the market like underreaction, overreaction, momentum, etc. as well as extreme movement in the market. Models of investor sentiment in subsections 2.3.3 and 2.3.4 explained the role of investor sentiment in market phenomena like underreaction, overreaction and momentum. Some of the empirical evidences related to them are reported below:

Thaler and De Bondt (1985) gave empirical evidence for the presence of overreaction over long periods (3 or more years) by creating winner and loser portfolios each consisting of 35 stocks from NYSE and then computing cumulative average returns. They found that the Loser portfolios of 35 stocks outperform the market on an average by 19.6%, thirty-six months after portfolio formation. Winner portfolios, on the other hand, earn about 5.0% less than the market. So the difference in cumulative average residual between the extreme portfolios, equals 24.6% (t-statistic: 2.20).

Zarowin (1989) found that prior months extreme losers significantly outperform extreme winner of subsequent month irrespective of their size. Small losers are found to earn greatest abnormal returns in January. This provides an evidence for short-run overreaction (1 month).

Dhankar and Maheshwari (2014) provided an evidence for a strong short term momentum effect and long term overreaction effect in Indian stock market. With a short-term formation-holding period of 3 to 12 months, returns showed momentum effect giving high abnormal profits of 7.7%. They also show strong reversal effects giving a contrarian strategy abnormal returns of 35.7% for long term formation-holding period of 36 months. This also means focusing purely on the past price information, Indian investors can earn abnormal returns in the Indian stock market contrary to Efficient Market Hypothesis.

Barber and Odean (2008) propose a model of decision making wherein they find that unlike institutional investors, when individual investors are subjected to decision of selecting stocks from a large pool, they primarily pick up those stocks that have attention attractive qualities. However, the same is not true when selling stocks; the reason being that they tend to hold few stocks, thus having limited decision choices to make. They list three attention attractive ways which are followed by individual investors. These are stocks abnormal daily trading volume, the stocks (previous) 1-day return, and whether the firm appeared in that days news. One implication of attention based buying by many investors is that stock price might go up for a brief period followed by disappointing subsequent returns thereby indicating overreaction.

Shefrin and Statman (1985) finds that investors at the time of positive news have a tendency to sell their stocks rapidly to realize gains before subsequent decline and during negative news they hold them expecting the price to recover. In other words, individual investors tend to sell winner stocks early and hold on to the loser stocks longer. They call this phenomenon as the disposition effect. Due to this conservative nature of disposition investors, prices never reach their intrinsic value indicating underreaction. To overcome this difference, rational investors generate momentum in stocks to push them towards their intrinsic value.

Overreaction models and evidences presented here show that there exists opportunities for generating abnormal returns if one creates a portfolio of stocks using rules based on these models contrary to EHM.

In summary, the behavioral finance models came into existence as their counterparts “the classical finance models” were too restrictive in their assumption of “a rational investor" making these models ineffective when stock prices deviated from their intrinsic value. Several studies discussed above felt the need to use the model of bounded rationality which relaxed the rationality assumptions. Doing so, they were able to explain market reaction using psychological and sociological biases in investors. The models of investor sentiment particularly focused on stock market underreaction, overreaction, momentum, etc.

2.4 News sentiment

The extant literature investigated the role of investors in stock market movement and found that their beliefs are important for them to take positions in the stock market. Investors reinforce their beliefs if the outcome of their investment match their existing beliefs and soften them if the outcome of their investment don’t match their existing beliefs. In this context, some researchers like Black (1986), De Long et al. (2003), Shleifer and Summers (1990), etc. further investigated the role of information on the beliefs of investors and revealed that information is often disguised with noise signals called sentiments that investors fail to decipher and often trade on sentiment thinking that it is information. Some investors called noise traders are highly susceptible to these sentiments because of their eagerness to trade. Since these sentiments are generated as a result of investor’s belief interacting with the information they receive, researchers felt that it is important to examine the content of information so as to find out what type of content creates a sentiment having significant influence on stock market. A pioneering study on the influence of media pessimism on US stock market variables like price and trading volume conducted by Tetlock (2007) using Wall Street Journals(WSJs) Abreast of the Market columns revealed that high levels of pessimism in the Wall Street columns predict downward pressure on market prices, followed by a reversion to fundamentals. Also, high or low values of media pessimism forecast high market trading volume. Conversely, low market returns lead to high media pessimism. This study presented a new dimension in the field of behavioral finance suggesting that media has a role to play in the changes occurring in the investor sentiment.

Findings of the research by Tetlock (2007), encouraged other researchers around the globe to probe different kinds of media and their influence on stock market. Recent literature on this strand of research suggests that online news media and social media are the favorites among the researchers because of their popularity and ease of processing their content available in digital form. Researchers like Ferguson et al. (2011), Ranco et al. (2015), W. Zhang and Skiena (2010), etc. used sentiment derived from online media useful to explain the behavior of stock price movement. One of the most common techniques used here is the sentiment analysis which is a method to analyze content of a piece of text to find out the affective component in it. Sentiment analysis is well supported by other related computing technologies like Natural Language Processing (NLP), Web Mining, News analytics, etc. which has made it possible to derive sentiment from unstructured text available in news and social media.

2.4.1 Definition of News sentiment

Early research in behavioral finance widely uses the term "investor sentiment" to mean pessimism and optimism among the investors. However, in the recent research, media sentiment has gained a lot of attention across different domains like marketing, politics, finance, etc. Although, extant literature doesn't seem to provide a legitimate definition for media sentiment, it indicates the sentiment prevailing in the content of the media. The term "media sentiment" is a generic term that is used to refer sentiment derived from any type of media including social media and news media. Chari et al. (2017) defined media sentiment as an implicit mood contained in the content of media which has a potential to change the investor psychology about a stock at a given point in time. Past research has used online sources like twits, blogs, news feeds, etc. along with certain columns published in finance journals. When sentiment is derived from news, it is called as the news sentiment. Referring to Chari et al. (2017), news sentiment can be defined as an implicit mood contained in the content of news which has a potential to change the investor psychology about a stock at a given point in time. Cahana (2013), Ferguson et al. (2011), Kräussl and Mirgorodskaya (2013) and many others who tried to find out the impact of news on stock prices, considers tone of news as media sentiment. In this research, focus is on the sentiment generated by online news published by a news portal, so the term "news sentiment" is used throughout this research.

2.4.2 Sentiment Analysis

Sentiment analysis is a type of text mining technique that discerns emotions and subjective opinion of a document or a group of documents. In recent times, it is widely used across different domains - in marketing for product reviews, in politics to reveal inclination of voters towards a party, in stock market to gauge the sentiment impact of a certain event on a stock(s), etc. One may use sentiment analysis for sentiment extraction, sentiment classification, subjectivity classification, summarization of opinions or opinion spam detection, etc.

Although computational treatment of unstructured data for sentiment classification and subjectivity classification can be found in behavioral finance literature for quite some time, advent of text mining, web mining, NLP and other tools made processing of unstructured data in text form easier. Pang and Lee (2008) have referred such computational treatment of text in the literature as sentiment analysis, opinion mining and subjectivity analysis.

Past research on stock markets mainly focused on specific information in quantitative form mostly about company's performance, economic indicators or market data even though narrations were available in the form of press releases and announcements. So, there is already abundance of knowledge created using quantitative form of data. It is important to augment the existing knowledge obtained from quantitative data with new findings available using unstructured text content.

Today, a lot of text content is generated on social media platforms like twitter, blog, facebook, etc. as well as news media like newspapers, news magazines, online newspapers, news blogs, rss (rich site summary) news feeds and news portals. Social media content in the form of blogs and twits generally are highly focused on a particular subject or entity. This makes their analysis a bit easier with respect to this characteristic, although they have other problems related to style of writing that includes emoticons and hashtags. News on the other hand, cover a wide spectrum of news articles across different categories which requires a proper collection and filtering mechanism to ensure that these news articles are related to the problem being researched. Online news sources, being digital in nature makes collection and filtering mechanism easier and convenient.

According to Medhat et al. (2014), sentiment analysis can be performed at three levels as given below:

1. Document-level: In document level sentiment analysis, the entire document is taken as a basic information unit talking about one topic and it is then classified as having positive or negative sentiment.

2. Sentence-level: Sentence level sentiment analysis classifies sentiment expressed in each sentence.

3. Aspect-level: Aspect levels sentiment analysis classify the sentiment with respect to the specific aspects of entities. Here, sentiment analysis process identifies entities and their aspect. An entity may give different sentiment for different aspects conveyed by the same entity. For instance, in a sentence "This Government wins elections but disappoints with their policies". Here Government is an entity which has two aspects in a sentence viz. election and policies with differing sentiments. While winning an election expresses positive sentiment, disappointment with policies expresses negative sentiment.

One of the most important elements of sentiment analysis process is the feature set used for sentiment classification. A feature set is a set of properties of the text that can be used in the sentiment analysis process for improving text classification. Giachanou and Crestani (2016) have listed four types of commonly used features in sentiment analysis:

1. Semantic Features: These include opinion words, sentiment words, semantic concepts and negation. Opinion words are words or phrases that convey some opinion about something. For example, you may ask a friend about his opinion on whether it is going to rain today? or Is Goa the best tourist destination spot in India? Sentiment words are those words that convey positive or negative tone in a sentence. For instance, a sentence "10 people killed in a bomb blast". Here the word "kill" gives a negative impression of the event. Semantic concepts refer to a coherent entity in the mind which can be represented by a cluster of symbols. For example, A toy block could be called a block, a cube, a toy. Negation refers to those words that flips the polarity of text. For example, the word "not" in the sentence "This is not good time to invest in the stock market" flips the meaning of the entire sentence to bad.

2. Syntactic Features: It includes features such as unigrams, bigrams, n-grams, terms frequencies, POS (Parts of Speech), dependency trees, and coreference resolution. Unigrams, bigrams and n-grams are often used for finding their frequency of appearance or presence in a given text. Unigrams refer to a single word, while bigrams and ngrams refer to collection of two and n words respectively. While "Bag of Words" doesn't care about position of words in a sentence, other methods do consider position of particular words in a sentence as an important feature. Natural Language Processing (NLP) based methods typically look for mostly adjectives and nouns along with verbs and adverbs in a sentence. In dependency trees, importance is given to linkages between various parts of speech so the words and other linguistic units are connected to each other by directed links. Coreference resolution refers to cases when two or more expressions refer to the same person or thing.

3. Stylistic Features: These features include emoticons, intensifiers, abbreviations, slang terms, and punctuation marks which are often found in non-standard writing like twits.

4. Twitter-Specific Features: These features include hashtags, retweets, replies, mentions, usernames, followers, and URLs.

Sentiment Analysis approaches

According to Medhat et al. (2014), sentiment analysis approaches can be broadly divided into Machine learning approach and Lexicon based approach. However, hybrid approach combining these two has been used in Malandrakis et al. (2013), Appel et al. (2017).

Machine Learning Approach: Machine learning approach involves the use of one or more algorithms to train the dataset and then use the trained classifier for testing. One of the most important tasks in machine learning approach is feature selection. Some of the commonly used features in machine learning include term presence and frequency, parts of speech (POS), opinion words and phrases, negations, etc. Machine learning approach can be further classified into supervised learning or unsupervised learning.

(i) Supervised Learning: Supervised learning approach involves building classifiers from labeled instances of texts or sentences. In supervised learning, training data is mostly manually labeled or pre-compiled by someone else. This trained classifier is then subjected for cross validation to check for high variance so as to ensure its relevance to unseen data. It is then used on test data. One can use various supervised learning methods, compare them and then select the one which gives best results. Evaluation metrics often employed to compare includes accuracy, precision, recall and F-score. They rely on the ability of the classifier to identify the cases correctly and can have the binary outcomes as shown in the confusion matrix in Table 2.1. Here true positive (TP) and true negative (TN) indicates that the predicted outcome is same as actual outcome. A false positive indicates that the predicted outcome was negative instead of the actual positive outcome. In the same manner, a false negative indicates that predicted outcome was positive instead of the actual negative outcome. Giachanou and Crestani (2016) have provided some important performance

Table 2.1: Confusion Matrix

	Predicted as Positive	Predicted as Negative
Are Positive	True Positive (TP)	False Negative (FN)
Are Negative	False Positive (FP)	True Negative(TN)

measurement parameters useful in sentiment analysis using these outcomes as given below:

Accuracy: Proportion of correctly labelled cases from all cases.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.7}$$

Precision: Also known as exactness, measures proportion of number of true positives from cases which were predicted as positives.

$$Precision = \frac{TP}{TP+FP} \tag{2.8}$$

Recall: Also known as sensitivity, measures proportion of true positives from cases that are actually positive.

$$Recall = \frac{TP}{TP+FN} \tag{2.9}$$

F-score also known as F1-score is a harmonic mean of recall and precision.

$$= 2 * \frac{Precision \times Recall}{Precision + Recall} \tag{2.10}$$

There are many supervised learning algorithms available for sentiment analysis, but Naive Bayes (NB) classifier, Support Vector Machine (SVM), Maximum Entropy (EM) are the most commonly used algorithms in research studies. Using Pang et al. (2002), let $f_1, ..., f_m$be a predefined set of m features that can appear in a document and let $n_i(d)$ be the number of times f_i occurs in document d. Then each document d is represented by the document vector $\vec{d} := (n_1(d), n_2(d), ..., n_m(d))$. Using these set of features, above three mentioned classifiers can be mathematically expressed as follows:

(a) Naive Bayes (NB) classifier: Naive Bayes classifier is a simple probabilistic classifier based on Bayes' rule given as

$$P(c|d) = \frac{P(c).P(d|c)}{P(d)} \tag{2.11}$$

where $P(c|d)$ is the posterior probability of class c given predictor document d, $P(c)$ is the prior probability of class c, $P(d|c)$ is the likelihood which is the probability of predictor document d given class c, $P(d)$ is the prior probability of predictor document d. In Naive Bayes classifier, a class c^* is identified based on $c^* = argmaxP(c|d)$. and assumes that the features f_i's are conditionally independent given d's class. Using features in a document, Bayes rule can be further expressed as follows:

$$P(c|d) = \frac{P(c).(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)}}{P(d)} \tag{2.12}$$

(b) Maximum Entropy (EM): Unlike NB classifier where each feature is conditionally independent given it's document, the maximum entropy classifier searches weights for features depending on the likelihood of training data. The basic principle behind maximum entropy is that the distribution should be as uniform as possible when when nothing is known. In other words, we call it as maximal entropy. According to Nigam et al. (1999) $P(c|d)$ for Maximum Entropy can be written as:

$$P_{ME}(c|d) = \frac{1}{Z_d} \exp(\sum_i \lambda_i F_i(d,c)) \tag{2.13}$$

where Z(d) is a normalization function. F_i is a feature/class function for feature f_i and class c, defined as follows:

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \quad \text{and } \acute{c} = c \\ 0 & otherwise \end{cases} \tag{2.14}$$

where each $f_i(d,c)$ is a feature, λ_i is a parameter to be estimated and $Z(d)$ is a normalizing factor to ensure a proper probability given as:

$$Z(d) = \sum_c \exp(\sum_i \lambda_i F_i(d,c)) \tag{2.15}$$

(c) Support Vector Machine (SVM): SVM is a non-probabilistic classifier suitable for text classification because of sparse nature of text. The main principle behind SVMs is to determine linear separators in the search space which can best separate the different classes. In the case of text classification between two classes, training procedure identifies a hyperplane represented by vector $\vec{w}$ such that it separates the document vectors into these two classes with maximum separation or margin possible from each class data points. In a set of two classes positive and negative represented as +1 and -1, if $c_j \in \{+1,-1\}$ is the correct class of document d_j, the hyperplane $\vec{w}$ can be defined as:

$$\vec{w}_j = \sum \alpha_j c_j \vec{d}_j, \quad \alpha_j > 0 \tag{2.16}$$

where α_j's are obtained by solving the dual optimization problem. Those $\vec{d}_j$'s whose α_j's are greater than 0 are called support vectors as they are the only document vectors contributing to $\vec{w}j$

(ii) Unsupervised Learning: Unsupervised learning on the other hand is useful when labeled training data is not available and it takes a lot of time to label a corpus of significant size. After labelling the corpus, it can be used further for supervised learning (Boiy et al., 2007).

Lexicon Based Approach: In lexicon-based approach, sentiment of a piece of text is obtained by checking presence of certain words/phrases of document as listed in a sentiment lexicon. Augustyniak et al. (2016) have defined sentiment lexicon as a set of ngrams (one or more consecutive words) with sentiment orientation assigned to these ngrams. Kolchyna et al. (2015) have described a typical process involved in sentiment analysis using lexicon based approach. Three steps involved in this process are as follows:

1. Preprocessing: Typically sentiment analysis process begins with collection of text data into a repository called text corpus. Once the text is obtained in the corpus, it has to be pre-processed to make it suitable for the sentiment extraction method. Preprocessing also

improves classification and reduce the computational complexity. Some of the steps often used in lexicon based sentiment analysis preprocessing step are given below:

- Part of speech (POS) tagging: Here each word in the corpus is tagged according to which part of speech it belongs to: noun, pronoun, adverb, adjective, verb, interjection, intensifier, etc. Pak and Paroubek (2010) have shown how the POS tagging can be used as an indicator for positive and negative text. For instance, superlative adverbs like “most” and “best” indicate positive text while verbs in the past tense like “missed”, “bored”, “gone”, “stuck”, “lost”, “taken”, etc. indicate negative text.
- Stemming and lemmetization: Stemming is a process where in words are brought to their roots or stems by trimming the ends of words using a crude heuristics with an objective of reducing inflectional forms and sometimes derivationally related forms of a word to a common base form. For example, words like “speaker”, “speaking” are brought to their root word “speak” after stemming. Stemming increases recall at the time of indexing and searching. Jivani (2007) have compared different types of stemming algorithms like Lovins Stemmer, Porter stemmer, Lancaster stemmer, N-gram Stemmer, Krovetz Stemmer (KSTEM), etc. Of these stemmers, Porter stemming is one of the oldest and the most popular one. Lemmatization refers to getting the words to their base or dictionary form of a word, known as the lemma by removing inflectional endings only. This is done by using vocabulary and morphological analysis of words. For example, a word “studies” after lemmatization will return “study” while in stemming the word returned would be “studi”.
- Stop-words removal: Stop words are the most common words in text which mostly does connection function in a sentence such as prepositions and articles. “a”, “an”, “the”, “at”,“which” are some of the stopwords we often see in sentences. These words need to be removed as they do not have any sentiment value in a sentence.
- Contextual valence shifters: The degree of negativity or positivity in a sentence can be varied by use of certain words in the sentence. The term “valence” refers to positive or negative attitude words found in sentences. At the basic level we have negation words or modifiers which convert the sentiment of a piece of text from positive to negative or vice versa. Some of the examples of negation words are “no”, “not”, “don’t”, “hardly”, “cannot”, “wasn’t”, “haven’t”, etc. For example: clever (+2) combined with not gives not clever (-2). Since, negation words completely change the polarity of a sentence, one needs to handle such negation words carefully.

  Intensifiers are the words that strengthens or weakens the base valence word. For example: Adding deeply to the word suspicious (-2) leads to deeply suspicious (-3). Here the word “deeply” strengthens the base valence word “suspicious” from (-2) to

(-3). Similarly, adding the word “rather” to efficient (+2) leads to rather efficient (+1). Here the word “rather” weakens the base valence word “efficient” from (+2) to (+1).

Connectors such as “although’, “however”,“ but”, “on the contrary”, “notwithstanding”, etc. generally change the polarity of the part of the sentence following them. In the sentence “Although Raj is brilliant with his batting skills, he is a horrible fielder”. Without the word “although”, the two valence words “brilliant” (+2) and “horrible” (-2) leads to a neutral sentence (0), however presence of the word “although” neutralizes the positivity in “brilliant” with final result being a negative sentence (-2).

A detailed discussion on the contextual valence shifters stated here as well as many others like modals, irony, presuppositional items, etc. is available in Polanyi and Zaenen (2006).

- Tokenisations into N-grams: Tokenization is a process by which an incoming string of words is broken down into individual words (unigram) or collection of two neighbouring words (bigram), collection of three neighbouring words together (trigrams), etc. N-gram tokenization can decrease bias, but because of multiple combinations of words may increase statistical sparseness.

2. Determination of sentiment polarity of each word in BOW: After preprocessing if one uses an approach called Bag of Words (BOW), preprocessed text is available as repository of terms (words) without any order. In BOW approach each word is checked against words in sentiment lexicon to determine its polarity. If the word from BOW is not found in the sentiment lexicon, it gets the polarity as zero.

3. Sentiment score of each document: Once all the words in BOW are assigned with their respective polarities, different measures can be used to obtain sentiment of each document in the corpus; some of which are given below:

   - Ratio of negative words: Ratio of words labelled as negative in that dictionary compared to the total number of words in the document

$$R_n = \frac{N}{T} \tag{2.17}$$

   - Ratio of positive words: Ratio of words labelled as positive in that dictionary compared to the total number of words in the document.

$$R_p = \frac{P}{T} \tag{2.18}$$

   - Sentiment ratio: Ratio of words in that dictionary compared to the total number of words in the document.

- Sentiment score: Difference between positive and negative word counts divided by the total number of words. Here sentiment score of a document lies between -1 to +1. If the sentiment score equals zero, then document is considered to be having neutral sentiment polarity. If there are P number of positive words, N number of negative words and T denotes total number of words, sentiment score S is given as:

$$S = \frac{P - N}{T} \tag{2.19}$$

- Sentiment polarity score: It is the difference between positive and negative word counts divided by the sum of positive and negative words. Here sentiment polarity score of a document lies between -1 to +1. If there are P number of positive words, N number of negative words, sentiment polarity score SP can be calculated:

$$SP = \frac{P - N}{P + N} \tag{2.20}$$

According to Guo et al. (2016), Bag of words (BOW) is one of the simplest and most popular approaches widely used by finance and accounting researchers . In this method, sentiment of each word is obtained as negative or positive and then aggregated to get overall sentiment of piece of text. Order of words or their relationship with one another is not considered.
Lexicon-based approach is further divided into dictionary-based approach and corpus-based approach.

1. Dictionary-based approach: Dictionary-based approach uses predefined dictionary of positive and negative words, and then counts words with positive and negative valence in the review text by matching them against words in the dictionary. Some popular dictionaries often used for research using lexicon based approach in sentiment analysis include MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon (Wiebe et al., 2005), Harvard General Inquirer (Stone & Hunt, 1963), Loughran-McDonald financial sentiment dictionary (LM)(Loughran & Mcdonald, 2011), Henry's financial sentiment dictionary (HE) (Henry, 2008), etc. Dictionaries like Sentiwordnet, SentiWordNet, WordNet, SentiTFIDF, SentiFul, SenticNet, etc. makes it possible to search opinion seed words in the dictionary for their synonyms and antonyms. Medhat et al. (2014) has explained how this process works using Wordnet dictionary: The process starts with a small set of opinion words collected manually with known orientations which is grown by searching for their synonyms and antonyms. The new found words goes through the same process of searching their synonyms and antonyms till no new word is found. Once this process is completed, manual inspection can be done to eliminate or correct errors.

2. Corpus based approach: Corpus-based approach is useful when one needs a context specific list of sentiment oriented words and it begins with a seed list of opinion words, and then finds other opinion words in a large corpus. This could be done by using statistical or semantic methods (Medhat et al., 2014).

2.5 News Sentiment and related research

Ever since Tetlock (2007) documented that sentiment in the news predict future stock market returns using "Abreast of the Market" columns from Wall Street Journals by using General Inquirer, there has been a flurry of research investigating relationship between different types of media and the stock market. News media has been one of the most popular media being investigated for its influence on stock market by researchers like Chowdhury et al. (2014), Ferguson et al. (2011), W. Zhang and Skiena (2010), etc. One of the forms in which news is available is text which is unstructured unlike the numeric data. One needs to identify relevant variables from it so as to examine relationship between stock market variables and news variables. In the context of volatility in stocks, Shi et al. (2016) pointed out that the past literature provides an explanation for relation between public information arrival and volatility under two streams viz. rate of information arrival and news sentiment. Various researchers across the globe have conducted studies using news sentiment and different variables related to stocks. A review of their work in the context of this thesis is presented here under following heads:

- Methodologies followed to test influence of news sentiment
- Duration of influence of news sentiment on stocks
- Sentiment analysis using other media sources
- Spillover effect of news sentiment

2.5.1 Methodologies followed to test influence of news sentiment

Schumaker and Chen (2006) used Bag of Words, Noun Phrases, Proper Noun and Named Entities as textual representations in Support Vector Machine algorithm for analyzing financial news articles and examined predictability of returns in a window of 20 minutes after "breaking news" release. They compared various combinations of textual analysis techniques and found which one is most suitable for stock price prediction. They used four models based on Support Vector Machine for comparison but varied the data fed to each one. These four models included: (1) Regress: Simple linear regression estimate of the +20 minute stock price (2) Model M1: with only article terms for its prediction (3) Model M2: with article terms and the stock price at the time the article was released and (4) Model M3: with article terms and a regressed estimate of the +20 minute stock price. They found Model 2 as the best performer in all three metrics; measures of Closeness at 0.04261, Directional Accuracy at 57.1% and Simulated Trading at a 2.06% return. Proper Nouns had the better textual representation performance on Directional Accuracy at 58.2% and Simulated Trading at 2.84% compared to the rest while it suffered on poor Closeness measures. Because of noisy terms in the articles, Bag of Words performed poorly by comparison.

Drury et al. (2012) highlighted the flaw of dependency on human annotator in models for sentiment classification based on manually selected training data or manually constructed dictionaries as well as the approach of manually aligning news stories with trends in a specific market. They designed a news classification strategy by combining rule classifier, alignment and self-training and showed that this strategy is better than competing models by evaluating them on F-measure and estimated trading returns. For testing this strategy, they collected more than 300,000 stories from a variety of news sites via Really Simple Syndication (RSS) feeds during the period from October 2008 to June 2010 along with FTSE stock market data. Also, they provided a comparative evaluation of news stories using different classification methods headline, description or story text with Language Models and Naive Bayes classifier.

Nagar and Hahsler (2012) built a news engine for gathering and aggregating news items from Google Finance News Feeds in real time and analyzed them for key financial terms and phrases to obtain their sentiment polarity and score. Filtering on news was done based on the ticker symbols of stocks and only the relevant part of the news article related to a particular stock was considered for analysis. They created a database of sentiment words using R package “tm.plugin.tags”. They used Natural language Processing tools and Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon to analyze text. They found that news sentiment time series generated for stocks shows a very strong correlation with the actual stock price movement.

Chowdhury et al. (2014) used real time news headlines and press-releases from large number of news sources like Wall Street Journal, Financial Times, Forbes, Reuters, CNBC, NDTV, Economic Times, Hindustan Times, Times of India, India Times, Telegraph, CNN Money, Market Watch, and Fortune to create a predictive model based on news sentiment after filtering relevant news. The classifier in this predictive model was designed using dictionary based approach which classified text into positive, negative and neutral categories. They analyzed this model on following parameters: Precision, Recall, Specificity, False Discovery Rate, Accuracy, F1 harmonic factor, Mathews Correlation Coefficient, Error in prediction, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Coefficient of Variation of RMSE. They found that their model is able to get average accuracy score for identifying correct sentiment of around 70.1%, RMSE of 30.3%, MAE of 30.04% and an enhanced F1 factor of 71 percent. They predicted the sentiment of 15 companies over a period of 4 weeks and showed that there is 67 % correlation between positive sentiment curve and stock price trend.

Mao et al. (2011) compared predictability of Dow Jones Industrial Average, trading volumes, and market volatility (VIX), as well as gold prices using six sentiment indicators- Daily sentiment

Index (DSI) bullish percentage, Investor Intelligence (II), Twitter Investor Sentiment (TIS), Tweet volumes of financial search terms (TV-FST), Negative News Sentiment (NNS) and Google search volumes of financial search terms called Google Insights for Search (GIS). For weekly analysis, they found significant correlation between weekly GIS and DJIA closing prices, trading volume and VIX. For daily analysis, they found all sentiment indicators having significant correlation with log returns and VIX. Similarly, TIS and the TV-FST values of the previous 1-2 days were found to be statistically very significant predictors, NNIS less significant predictor while DSI not a significant predictor of daily market returns.

Leinweber and Sisk (2011) identified some attributes of news useful for sentiment analysis. These are given below:

- News intensity: Number of news items in a given period
- Relevance: To what level (from 0 to 100%) the new item is applicable to a stock
- Probability of a news item being positive, negative, or neutral in tone.
- Novelty: Number of links to previous related items by time; News item is considered to have novelty if there are no links to previously published news items.

They used these attributes to demonstrate how news signals can be exploited to produce alpha. They found that basic materials, cyclicals, financials, industrials, non-cyclicals, and technology sectors are more favourable to produce alpha on news signals. Using simulation and event study they also showed that news signals can produce alpha in excess of 10% per year.

2.5.2 Duration of influence of news sentiment on stocks

Gidofalvi (2001) demonstrated that stock price movements can be predicted within a window of influence (20 minutes before and 20 minutes after) around the news release. For this, he collected price data for 127 stocks in intervals of 10 minutes along with timestamp of news articles released during the period from 14$^{\text{th}}$ November 1999 to 11$^{\text{th}}$ February 2000. He labeled movement of each stock price time series as “up”, “down”, or (approximately) “unchanged” relative to the volatility of the stock and the change in a relevant index. He used these labels on news articles on training data to train the Naive Bayes classifier and then applied a classifier to classify the news articles in the test data set. He found that although, the predictive power of the classifier was low, a strong correlation between news articles and the behavior of stock prices was present within the window of influence.

Ferguson et al. (2011) examined the effect of news sentiment on stock returns for UK companies and tested their return predictability. They showed that measures of positive and negative news sentiment had significant relationship with the returns on the day of news release. Also negative news sentiment had significant relationship with stock returns on the day following the news release, but this relationship was weaker compared to the one on the day of news release. Positive news sentiment did not show any predictive power on the day following the news release. Also, they found that the amount of media coverage was strongly related to both positive and negative media sentiment; however this relationship was weaker in case of positive news sentiment as compared to negative news sentiment.

Antweiler and Frank (2004) examined whether internet messages move markets. They collected more than 1.5 million messages posted on Yahoo Finance and Raging Bull about the 45 companies in the Dow Jones Industrial Average and the Dow Jones Internet Index. Using Naive Bayes and Support Vector Machine algorithms they derived bullishness index which was then used to forecast market volatility, trading volume, and spreads. They found that stock messages help predict market volatility with a significant influence on stock returns but this influence is economically small. They also found a positive correlation between message posting and trading volume. While at daily frequency, effect of message posting activity predicted trading volume significantly, at 15-minute frequency opposite effect was significant wherein increase in trading volume resulted in more message posting.

Kräussl and Mirgorodskaya (2013) hypothesized that investor's perception about future prospect as well as their sentiment is influenced by the way in which a newspaper article describes a current financial market state or presents new financial information. Because of this, investors may form certain expectations and update their investment decisions that can have a direct influence on the performance of the financial market. They examined whether pessimistic news media sentiment decreases (increases) financial market returns (volatility) in the long-run. Using news collected by searching a predetermined set of keywords in the LexisNexis database for the three news sources viz. New York Times, the Wall Street Journal Abstracts, and the Financial Times, they constructed a monthly media sentiment indicator by taking the ratio of the number of newspaper articles with predetermined negative words to the number of newspaper articles with predetermined positive words in the headline and/or the lead paragraph. They analyzed the potential media sentiment impact on financial market returns using vector autoregressive (VAR) model and Grangar causality test. They formed these two VAR models as given by the following two equations:

$$Mrkt = \alpha_1 + \beta_{11} L_{24}(Mrkt) + \beta_{21} L_{24}(Sentt) + \beta_{31} Exog_t + \varepsilon_{t_1} \tag{2.21}$$

$$Volat = \alpha_2 + \beta_{12} L_{24}(Volat) + \beta_{22} L_{24}(Sentt) + \beta_{32} Exog_t + \varepsilon_{t_2} \tag{2.22}$$

where, $Mrkt$ represents the log rate of return of the Morgan Stanley Capital International (MSCI) World index; $L_{24}(x_t)$ is a lag operator that transforms the variable x_t into a row vector consisting of 24 lags of x_t; $Sentt$ is the log change of media sentiment indicator; $Volat$ is the squared demeaned residuals of the MSCI World returns; $Exog_t$ are exogenous variables such as size Small Minus Big (SMB), value High Minus Low (HML), momentum (MOM), and liquidity (LIQ) and other macroeconomic variables used to control for other potential anomalies that are not driven by the news media. α_j are estimated constants and β_{ij} are estimated VAR coefficient. They provided an evidence of long term impact of news media effects on the global economy and also found causal relation of news media sentiment on the global market return and global market volatility for 12 to 24 months in advance.

Uhl et al. (2015) combined company and macro-specific news sentiment from around 1,00,000 news items per week published in Thomson Reuters global news universe and after removing noisy news sentiment using the CUSUM (cumulative sum) filter method, calculated momentum in news sentiment. They created a tactical asset allocation strategy using this news sentiment and found that this strategy outperformed other strategies with an information ratio of 0.8, while maintaining switches as few as eight times per year (on average). They found that there exists long term sentiment cycles which can be exploited with investment strategies based on news sentiment momentum. They suggested that this method can be used by both tactical asset allocators and investors in general.

Birz and Dutta (2016) studied impact of U.S. macroeconomic news on stock returns of 15 countries viz. Argentina (ARG), Brazil (BRA), Canada (CAN), Mexico (MEX), Austria (AUS), Denmark (DEN), France (FR), Germany (GER), Italy (ITA), Norway (NOR), Poland (POL), Spain (SPA), Sweden (SWE), United Kingdom (UK), and Japan (JPN). Using LexisNexis database, they collected newspaper articles from 389 newspapers resulting in a sample consisting of 12,620 headlines. A team of research assistants analyzed and classified headings of these articles into positive, negative, neutral, or mixed. They created a news index using the following formula:

$$\text{News index} = \frac{(\%Positive - \%Negative + 100)}{\sigma} \tag{2.23}$$

where standard deviation σ is used for normalization. A control variable S (for U.S. unemployment surprises) was used to show that newspaper coverage is a superior measure of economic news, i.e. it contains additional information that is not captured by surprises. In mathematical form, normalized value of this surprise element is given as follows:

$$S = \frac{A - F}{\sigma} \tag{2.24}$$

where A denotes the value of the released unemployment rate, F is the expected value of the unemployment rate provided by Money Market Services survey (MMS) forecasts, and σ is the sample standard deviation of $(A - F)$ used for normalization. They investigated whether months $(t-1)$ unemployment rate announced by Bureau of Labor Statistics Data (BLS) and covered by the media during month t has a statistically and economically significant effect on month t stock returns measured as a percentage change in the end-of-month index prices of month t and month $(t-1)$. They found that the U.S. unemployment rate news are important for 12 international stock markets out of 15 examined markets and that newspaper coverage is a superior measure of macroeconomic news as compared to macroeconomic surprises.

Heston and Sinha (2016) used dataset of news sentiment called Thomson Reuters NewScope of 9,00,754 articles tagged with firm identifiers from the Thomson-Reuters news system over the calendar years 2003 to 2010 and found that daily news predict stock returns for only one to two days and weekly news predict stock returns for one quarter. Positive news stories were found to increase stock returns quickly whereas negative stories were found to create a long delayed reaction. Also, firms with neutral news were found to outperform firms without any news.

2.5.3 Sentiment analysis using variety of media sources

Ranco et al. (2015) examined the relation between stock price returns and the sentiment expressed in financial tweets posted on Twitter. They obtained peaks of twitter volume and built a 15 month time series of sentiment of about 30 stock companies of the Dow Jones Industrial Average (DJIA) from tweets for each company using SVM (Support Vector Machine) classifier. They created a time series of volume of twits, negative tweets, positive tweets, neutral twits and sentiment polarity for each company in the sample. They examined causality and relationship of twitter sentiment with abnormal returns. Although they found relatively low Pearson correlation and Granger causality, significant dependence between the Twitter sentiment and abnormal returns was found during the peaks of Twitter volume.

W. Zhang and Skiena (2010) showed how a companys reported media frequency, sentiment polarity and subjectivity anticipates or reflects its stock trading volumes and financial returns. Here, they compared four different collections of text data comprising of dailies (news source) and three blog sources twitter, Spinn3r RSS Feeds and LiveJournal. To conduct this study they obtained polarity and subjectivity of the text content of various sources of text data using the equations given below:

$$polarity = \frac{p-n}{p+n}$$
$$subjectivity = \frac{p+n}{N}$$

where p and n denote the number of raw positive and negative references respectively and N is the total number of references in the corpus (including neutral references).

They found significant correlation between stock returns and data from these four sources in terms of reference frequency (article count), sentiment polarity and subjectivity. Other notable findings were: (1) Strength of correlation between logged normalized article counts and logged stock trading volume were found to be more than 0.4 (2) For Dailies, one day lag was having highest correlation (0.74) but persisted (between 0.64 and 0.68) for periods up to ten days in the future because of autocorrelation. (3) Trading volumes of Electronic & Electrical Equipments and Software & Computer Services sectors was less sensitive to media exposure (correlation < 0.2) as compared to Pharmaceuticals & Biotechnology and Aerospace & Defense sectors (correlation $>$ 0.7). (4) Correlation coefficients between article counts and stock trading volume was stronger for high market capitalization stocks compared to the ones with low market capitalization. (5) High positive correlation (0.42) between logged monthly normalized article count and logged market capitalizations. (6) Polarity and stock return pair was the most significant correlation among the six combinations of news variables (polarity, change in polarity and percentage change in polarity) and stock variables (change of stock price, stock return and stock abnormal return). (7) For news data, returns of day 0 was found to be having the best correlation with polarity and that today's news almost had no predictive power for returns on the next day as well the subsequent days. (8) News information took much shorter time period (within 1 day) to get incorporated into stock market completely as compared to blogs. They found this to be consistent with the efficient market hypothesis.

2.5.4 Spillover effect of news sentiment

Audrino and Tetereva (2017) studied cross-industry influence of the news for a set of US and European stocks. They found sentiment spillover effects dominating the direct effects when the market is driven by news coming from a different sector. They also found the strength of this effect growing just before periods of economic and financial instability and reaching a maximum during crises.

Mo et al. (2016) analyzed more than 12 millions of news articles and reports obtained from diverse group of media sources and using SentiWordNet dictionary obtained daily sentiment over a period of 660 days. They used this news sentiment and examined relationship with return time series of four exchange traded funds viz. SPDR[2] S&P 500, SPDR Dow Jones Industrial Average, PowerShares QQQ Trust[3] and SPDR Russell 3000. They found that there exists a significant feedback effect between news sentiment and market returns across these four major indices in the US financial market. Using the ordinary least squares method, they found strong relationship between news sentiment and S&P 500 market index returns as given by the equation:

$$R_t = 0.0026 + 0.0578 \times (News_t) \tag{2.25}$$

They also showed that news sentiment had a lag-5 effect on market returns and conversely market returns exhibited consistent lag-1 effects on news sentiment. With these results, they showed that news sentiment drives trading activity and investment decisions which are reflected in the returns

[2]Standard and Poor's Depository Receipt

[3]Now known as Invesco QQQ which is an exchange-traded fund (ETF) that tracks the Nasdaq 100 index

of these indices and also receives feedback from the market movement.

Smales (2016) investigated the empirical link between the news sentiment of specific news, implied volatility indices and stock index returns. This study was conducted using data over a period of 11 years from January 2000 to December 2010 obtained from Ravenpack that gives news sentiment in terms of Multi-classifier for Equities (MCQ). Implied volatility index (VIX) data and S&P500 index data were obtained from Thomson Reuters Tick History provided by Securities Industry Research Centre of Asia-Pacific (SIRCA). A significant negative relationship was found between changes in VIX and both news sentiment and stock returns in the same period. Also this relationship was found to be asymmetric wherein changes in VIX are larger following negative news and/or when stock market declines. Vector AutoRegressive (VAR) analysis revealed a strong positive relationship between previous and current period changes in implied volatility and stock returns. Current period and lagged news sentiment showed a significant positive (negative) relationship with stock returns (changes in VIX). Based on these findings, a trading strategy was suggested whereby investors can take a short (long) positions in the underlying index when there are high (low) levels of implied volatility. In the same manner, they need to take a short (long) position during extremely negative (positive) news sentiment signal. According to the findings of this paper, Investor fear gauge (VIX) was found to perform better than news sentiment measures in forecasting future returns.

2.6 Research on sectoral indices

Various types of studies have been conducted on sectoral indices in the past, some relevant ones in the context of this research are discussed below:

Guha et al. (2016) calculated volatility of NSE's eleven sectoral indices with respect to Nifty50 index and found that Realty, Metal and IT are the most sensitive to changes in Nifty50 index. They also found that FMCG, Pharma and Auto are the least sensitive sectors to Nifty50 index. Applying linear regression analysis, they also found that returns of Nifty50 index can be predicted to the tune of 95% using returns of six sectoral indices viz. Auto, Energy, FMCG, Metal, Pharma and PSU Banks. This study shows differences among various types of sectoral indices based on their volatility with respect to Nifty50 index. The characteristics of individual sectors can be utilized for creating strategies in stock market.

Aravind (2016) examined the short term and long term return trends of eleven sectors and analyzed momentum and contrarian strategies on them. They found an evidence of short term contrarian effect in sectors like Auto, Banking, Energy, Media and Metals sectors and momentum effect in sectors like Media, IT, Pharma, PSU, Realty and Finance. They also found an evidence of long term momentum effect in Auto, Bank, Finance, FMCG, IT, Energy, Metal, PSU, and Realty sectors and contrarian effect in Pharma and Media sectors.

Borovkova and Lammers (2017) developed the sector-based news sentiment indicators and showed their relationship with prices of a stock portfolios for companies in that sector (or the sector ETFs). They found that nearly for all sectors, the impact of news sentiment can be only seen during extreme returns i.e. high (in absolute value) positive or negative returns, but not at “normal" times. The magnitude of influence varied across different sectors with news sentiment showing most significant influence on the returns of financial sector as compared to other sectors. They also found that that a rising sentiment is accompanied by higher returns but surprisingly declining sentiment is also accompanied by higher returns and vice versa. Using their news sentiment indicator they developed simple sector rotation strategy to take a sell position in the sector when it showed declining sector-based sentiment.

Literature review conducted in this thesis reveals that early research in finance related to stock market movement was done mainly using quantitative data even though it was not the only type of data accessed by investors. In the past, text processing techniques were not advanced as compared to what they are today, so even though qualitative data was available it was difficult to process. Recent studies have shown that qualitative information is as important as quantitative data and has shown evidences of influence on stock market movement. Sentiment of company-specific news and its influence on stock's variables like price, returns, volatility, etc. has been one of the popular themes in recent research on stock market. However, considering the wide variety of events that happen in domestic and global market environment and their coverage in the news and social media, it gives a significant scope for researchers to examine sentiment induced by news or social media and find whether it influences stock market.

2.7 Research Gap and Research questions

2.7.1 Research Gap

Literature review conducted above reveals some gaps in research related to news sentiment and its influence on stock market as discussed below:

- Gap in literature regarding influence of news sentiment during extreme movement in stock market: Noise trader theory and other derived research from it points towards limits on arbitrage and sentiment as factors that give rise to large deviation in prices. While past literature reveals sentiment being more prominent during large deviations, influence of news sentiment during extreme movement is not well documented. For instance, W. Zhang and Skiena (2010), Ferguson et al. (2011), Chowdhury et al. (2014), etc. explain the response of stock variables like price, returns, volatility, etc. to stock-specific news sentiment and hence do not entirely pertain to extreme movement. So, there is a need for conducting a study that focuses on extreme movement in stock market and find whether there is any influence of news sentiment in this movement. This will also be an useful contribution towards extending the scope of noise trader theory by including influence of news sentiment as an important attribute having an impact on investor sentiment.

- Gap in literature regarding aggregate news sentiment and its influence on stock market: While Ferguson et al. (2011), Mo et al. (2016), Heston and Sinha (2016) and other similar studies reveal that sentiment obtained from company's news influences its returns, they do not take into account influence of other news released on the same day. This means that these studies are conducted by considering only company-specific news thereby undermining the influence of other news released during the same period. Everyday investors come across various types of news which may influence them to take positions in those stocks that are likely to be impacted by news. When various types of news are taken together, they will generate an aggregate news sentiment due to sentiment arising from news having same or differing polarity. On a given day, aggregate news sentiment may be such that it may have a higher probability of influencing investors.

- Gap in literature regarding influence of news sentiment from other news categories: While studies mentioned above have predominantly focused on company specific news sentiment and its influence on respective stock variables, there is not much evidence available in past research about sentiment of other categories of news and their influence on stock market. Event studies based on quantitative data by MacKinlay (1997), Quaye et al. (2016), Liew and Rowland (2016), Geetha et al. (2011), etc. suggest events related to business, politics, economy and international categories influence stock market returns. However, these studies are conducted only using quantitative data and not qualitative data. There is a scope for examining news sentiment related to these categories for their influence on stock market.

- Gap in literature regarding influence of news sentiment on portfolio of stocks: While a large number of studies like Cahana (2013), Ferguson et al. (2011), Heston and Sinha (2016), Huynh and Smith (2017), Mo et al. (2016), W. Zhang and Skiena (2010), etc. have focused on studying influence of company specific news sentiment on individual stocks, there is not enough documentation about studies examining relationship between news sentiment and portfolio of stocks. It would be worth examining the influence of news sentiment on different portfolios of stocks and find out if a certain type of portfolio is strongly influenced by news sentiment than others.

- Gap in literature regarding time-based spill-over effect: While Audrino and Tetereva (2017) provide some evidence for spill over effect of news sentiment across sectoral indices, there is not enough literature providing evidence for time-based spillover effect of news sentiment.

In summary, there is a potential to conduct a study on influence of news sentiment during extreme stock market movement, influence of aggregate news sentiment and category-wise news sentiment on stock market, influence of news sentiment on a portfolio of stocks and time-based spillover effect of news sentiment. In this thesis influence of aggregate news sentiment on returns of stock market as well as returns of select sectoral indices is examined. This study is conducted by having a 5-day observation window, with day of extreme market movement being in the middle of the window. Spillover effect of news sentiment of the first two days of the window is examined for their influence on the returns of stock market on the day of extreme market movement. News sentiment from business, economy, politics and international news categories are also examined for their influence on the returns of stock market as well as returns of select sectoral indices.

2.7.2 Research Questions

After conducting a thorough literature review followed by identification of gaps, the following research questions can be raised for further investigation:

- Whether there is any influence of aggregate news sentiment on the stock market returns during extreme movement in stock market[4]?
- Whether news sentiment related to business, economy, international and politics news categories influence stock market returns during extreme movement in stock market?
- Whether there is any spillover effect of aggregate news sentiment on stock market returns during extreme movement in stock market?
- Whether there is any influence of aggregate news sentiment on returns of various sectoral indices during extreme movement in stock market?
- Whether news sentiment related to business, economy, international and politics news categories influence returns of various sectoral indices during extreme movement in stock market?

[4]Stock market here refers to the Nifty50 index

2.8 Objectives

[5] The study conducted in this thesis focuses on the following objectives:

1. To study the relationship between aggregate news sentiment and stock market returns during extreme movement in stock market.
2. To examine whether there is any spillover effect of aggregate news sentiment on stock market returns during extreme movement in stock market.
3. To study the relationship between news sentiment of select news categories and stock market returns during extreme movement in stock market.
4. To study the relationship between aggregate news sentiment and returns of select sectoral indices during extreme movement in stock market.
5. To study the relationship between news sentiment of select news categories and returns of select sectoral indices during extreme movement in stock market.

2.9 Operational Definitions

Aggregate news sentiment: Aggregate news sentiment is the net news sentiment resulting from sentiment polarity of various news articles published in a given period having a potential to influence investors in the stock market.

Extreme returns: Extreme returns are defined as the Nifty50 index logarithmic returns beyond (+/-2.58) standard deviations of these returns.

Observation window: Observation window is a period of examining relationship between aggregate news sentiment and returns of stock market and consists of n consecutive trading days where n is an odd number and day $t_{(\frac{n+1}{2})}$ is the day of extreme stock market movement.

2.10 Hypotheses

Based on the above research objectives, the following hypotheses are framed:

2.10.1 Relationship between aggregate news sentiment and stock market returns during extreme movement in stock market.

Hypothesis 1 *There is a significant relationship between aggregate news sentiment and returns of the stock market on day (t_0)*

[5] Objectives stated earlier in chapter 1 are written again here for framing hypotheses

Hypothesis 2a *Relationship between aggregate news sentiment and stock market returns is stronger on day (t_0) as compared to day (t_{-2}).*

Hypothesis 2b *Relationship between aggregate news sentiment and stock market returns is stronger on day (t_0) as compared to day (t_{-1}).*

Hypothesis 2c *Relationship between aggregate news sentiment and stock market returns is stronger on day (t_0) as compared to day (t_{+1}).*

Hypothesis 2d *Relationship between aggregate news sentiment and stock market returns is stronger on day (t_0) as compared to day (t_{+2}).*

2.10.2 Spillover effect of aggregate news sentiment on stock market returns during extreme movement in stock market

Hypothesis 3a *There is a spillover effect of aggregate news sentiment of day (t_{-2}) on stock market returns of day (t_0).*

Hypothesis 3b *There is a spillover effect of aggregate news sentiment of day (t_{-1}) on stock market returns of day (t_0).*

2.10.3 Relationship between news sentiment of select news categories and stock market returns during extreme movement in stock market

Hypothesis 4 *There is a significant relationship between news sentiment of select news categories (business, politics, economy, and international) and stock market returns of day (t_0).*

Hypothesis 4a *News sentiment from the business category has a significant influence on stock market returns of day (t_0).*

Hypothesis 4b *News sentiment from the economy category has a significant influence on stock market returns of day (t_0).*

Hypothesis 4c *News sentiment from the international category has a significant influence on stock market returns of day (t_0).*

Hypothesis 4d *News sentiment from the politics category has a significant influence on stock market returns of day (t_0).*

Hypothesis 5a *Relationship between news sentiment of select news categories and stock market returns is stronger on day (t_0) as compared to day (t_{-2}).*

Hypothesis 5b *Relationship between news sentiment of select news categories and stock market returns is stronger on day (t_0) as compared to day (t_{-1}).*

Hypothesis 5c *Relationship between news sentiment of select news categories and stock market returns is stronger on day (t_0) as compared to day (t_{+1}).*

Hypothesis 5d *Relationship between news sentiment of select news categories and stock market returns is stronger on day (t_0) as compared to day (t_{+2}).*

2.10.4 Relationship between aggregate news sentiment and returns of select sectoral indices during extreme movement in stock market

Hypothesis 6 *There is a significant relationship between aggregate news sentiment and returns of select sectoral indices on day (t_0)*

Hypothesis 7a *Relationship between aggregate news sentiment and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{-2}).*

Hypothesis 7b *Relationship between aggregate news sentiment and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{-1}).*

Hypothesis 7c *Relationship between aggregate news sentiment and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{+1}).*

Hypothesis 7d *Relationship between aggregate news sentiment and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{+2}).*

2.10.5 Relationship between news sentiment of select news categories and returns of select sectoral indices during extreme movement in stock market

Hypothesis 8 *There is a significant relationship between news sentiment of select news categories (business, politics, economy, and international) and returns of select sectoral indices on day (t_0).*

Hypothesis 8a *News sentiment from the business category has a significant influence on returns of select sectoral indices on day (t_0).*

Hypothesis 8b *News sentiment from the economy category has a significant influence on returns of select sectoral indices on day (t_0).*

Hypothesis 8c *News sentiment from the international category has a significant influence on returns of select sectoral indices on day (t_0).*

Hypothesis 8d *News sentiment from the politics category has a significant influence on returns of select sectoral indices on day (t_0).*

Hypothesis 9a *Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{-2}).*

Hypothesis 9b *Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{-1}).*

Hypothesis 9c *Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on day (t_0) as compared to day (t_{+1}).*

Hypothesis 9d *Relationship between news sentiment of select news categories and stock market returns is stronger on day (t_0) as compared to day (t_{+2}).*

Chapter 3

Research Methodology

This chapter presents the method used to gather data, describes the econometric, statistical and text mining techniques used to answer research questions and provides details of the method used in this study. It is divided into the following sections:

1. Data collection
2. Econometric, Statistical and Text mining techniques
3. Method

3.1 Data collection

In this thesis, secondary data spanning over 12 years starting from 1st January 2008 to 31st December 2019 is used. This period of study is selected based on the following reasons pertaining to data:

- Recency: Recency is an important criteria because it is easier to comprehend current state of affairs through the data which is most recent. Also, any findings obtained for recent data has a higher probability in terms of its applications in the near future.

- Representativeness: Since the focus of this study is sentiment analysis of news, it is important to get data which represents the realistic setup within which a number of market participants take decisions even though study is undertaken using past data rather than in real time. Being the most recent data, market participants' must have had access to this information during that period using online media that ideally captures his/her decision making environment.

- Availability: Information available to market participants through online media has grown manifolds during the last ten years and many of these market participants must have had access to information using mobiles, laptops and desktop computers. In India, internet speeds have seen a considerable improvement and has a wider reach in the last ten years as compared to previous years.

- Quantity: To conduct a research study, one needs to have a good amount of data as a representative sample. Data available as archived news from the source selected offers a sizable amount of relevant data during this period.

- Quality: Quality of data in this study can be considered in terms of content that is suitable for sentiment analysis. Data available during this period is available as html pages from which relevant text can be obtained using web mining tools.

- Ease: The period of study selected here is such that relevant data required for both news sentiment and stock market variables is available easily. Also, news articles required for this thesis are available in the archive of the news portal selected in this thesis. So, data collection can be done with ease and without any cost.

As this study focuses on studying relationship between aggregate news sentiment and stock market, secondary data is used that comprises of the following:

1. Stock market data: Since the study is conducted for indian stock market, an ideal candidate to represent its behavior is the Nifty50 index. It is a diversified 50 stock index comprising of stocks from 13 sectors of the economy. It tracks the behavior of a portfolio of blue chip companies, the largest and most liquid indian securities capturing approximately 65% of its float- adjusted market capitalization. Daily adjusted closing prices of Nifty50 index for a 12 year period starting from 1^{st} January 2008 to 31^{st} December 2019 have been collected from India Index Services & Products Limited (IISL) website “https://www.niftyindices.com/”. IISL is a subsidiary of NSE Strategic Investment Corporation Limited. IISL was setup in May 1998 to provide a variety of indices and index related services and products for the Indian capital markets. Along with its widely popular portfolio Nifty50 index, it owns and manages a portfolio of 73 indices under Nifty brand (as of 31^{st} March 2017). Nifty indices are used as benchmarks for products traded on NSE and other venues. Plot of Nifty50 adjusted closing prices is shown in Figure 3.1

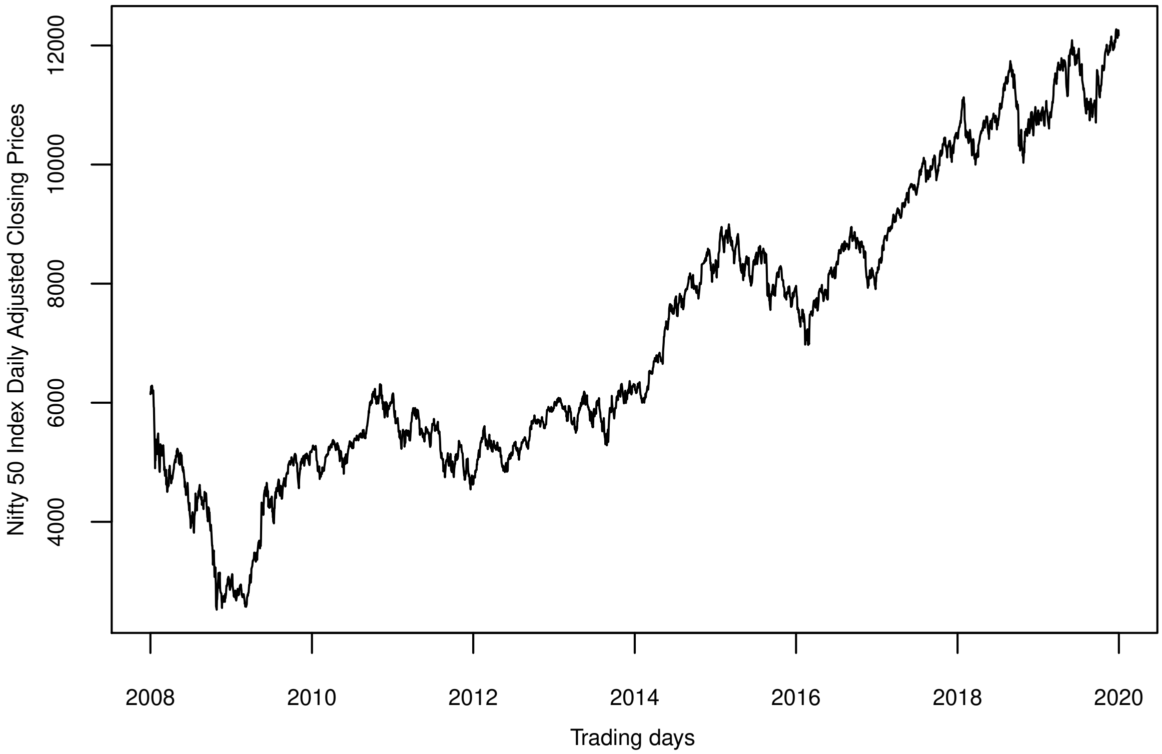


Figure 3.1: Nifty50 Index daily closing prices

In addition to the stock market data, adjusted closing prices of ten sectoral indices during the same period is also used in this thesis. These ten sectoral indices include Nifty Auto, Nifty Bank, Nifty Energy, Nifty FMCG, Nifty IT, Nifty Media, Nifty Pharma, Nifty Private Bank, Nifty PSU Bank and Nifty Realty.

2. News articles: This is an unstructured data that comprises of a collection of news articles released on specific days as determined by the methodology of the study but during the same period as the stock market and sectoral index data. It is important to gather news released from a source which is popular, authentic and having a good coverage of relevant news. Also, since events identified are past events, the source should have a good repository of archived news. In India, some of the popular web portals which provide freely available news archive include www.moneycontrol.com, financialexpress.com, economictimes.indiatimes.com, etc. In this research, news archive from economictimes.indiatimes.com is used as it fulfils the conditions stated above. It is an online news portal from Times group and is the worlds second-largest English business daily. Times group is also known for its daily The Times of India which is the largest English newspaper in the world. News archives of economic times are available on their online portal from the year 2001 onwards till date.

3.2 Econometric, Statistical and Text mining techniques

In order to conduct this study, the following econometric, statistical and text mining techniques are used:

3.2.1 Econometric technique: Autoregressive Integrated Moving Average (ARIMA)

Classical regression is inadequate to explain the time series that contains autoregressive and moving average components in it. If the time series contains only autoregressive components in it, Autoregressive (AR) models can be used. In AR models, future value of a variable is dependent on the past values of the variable. The generalized AR(p) model for the current value Y_t with p lag variables can be written as:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t \tag{3.1}$$

where c and ϕ_i are constants and ε_t is an error term at time t which we consider as white noise. Similarly, if the time series contains only moving average components in it, Moving Average (MA) models can be used. In MA models, future value of a variable is dependent on the past error terms of the time series. The generalized MA(q) model for the current value of Y_t with q lag error terms can be written as:

$$Y_t = d + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{3.2}$$

where d and θ_j are constants and ε_t is an error term at time t which we consider as white noise. A more complete model that includes both autoregressive as well as moving average terms is the combined ARMA model. The generalised ARMA(p,q) model for the current value of Y_t can be written as:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{3.3}$$

Generally lagged terms p and q should not be too large as this makes the coefficients too small to be statistically significant, may lead to overfitting thereby reducing predictive power and also interpretation becomes difficult with many terms. All these models AR(p), MA(q) and ARMA(p,q) can deal with a time series which is stationary. They cannot be used for non-stationary series. Box and Jenkin's Autoregressive Integrated Moving Average (ARIMA) model, also known as Box Jenkins methodology can be used to model a time series with non-stationarity. Given a non-stationary time series, it can be converted to stationary by a process called as "differencing". In differencing, difference between successive time lags in the time series is taken. This process stabilizes the mean of the time series by removing trends and seasonal components. For instance, in stock market Nifty50 index data might have trends over certain periods, but on a daily basis, change in Nifty50 index will likely to be centred at zero. If differencing once, does not give a stationary time series, one may repeat the process. However, repeating it more than twice may lead to reduction in explanatory power of the model. In addition to differencing, sometimes log transformation is also used to normalize variances so as to make the time series stationary. The generalized ARIMA(p,d,q) with p as autoregressive parameter, q as moving average parameters and d, the degree of differencing, time series modelled using ARIMA can be represented mathematically as follows:

$$Y_t^d = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{3.4}$$

ARIMA procedure : Box Jenkin's ARIMA methodology consists of the following steps:

1. Stationarity Check: A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Stationarity is important in a time series because most forecasting methods assume that a distribution has stationarity. For instance, autocorrelations and autocovariance rely on the assumption of stationarity. A time series with trends or seasonality (regular cycles) cannot be considered to be stationary. Absence of stationarity can lead to unexpected results, like t-ratios not following a t-distribution or high r-squared values assigned to variables that arent correlated at all. A time series which is not stationary can be made stationary through a widely used method called differencing where difference of lags is taken successively till the time series becomes stationary. This process stabilizes the mean thereby eliminating trend and seasonal components.
   The first step in ARIMA modelling is to check whether the given series is stationary or not. This can be done by using Augmented Dickey Fuller (ADF) Test. ADF tests the null hypothesis that a unit root is present in time series data. A unit root is a feature of a stochastic process that can cause problem in the statistical inference involving time series models. If the p-value is less than or equal to 0.05, it indicates rejection of null hypothesis. In other words the series is stationary. Other methods of checking stationarity include Phillips-Perron (PP) test, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. With these tests if the time series turns out to be non-stationary, then as stated earlier to make it stationary, differencing method is applied. The lowest order of differencing that makes the time series stationary is preferred.

2. Identification of the order of ARIMA model: Order of ARIMA model can be identified using autocorrelation function (ACF) and partial correlation function (PACF) plots which are also useful for checking stationarity. ACF plot is a bar chart of the coefficients of correlation between a time series and its lags. Similarly, The PACF plot is a plot of the partial correlation coefficients between the series and its lags. In ARIMA model, order terms include 'p' for

autoregressive term, ‘d’ for differencing term, ‘q’ for moving average term and the seasonal terms if the data series show seasonality. If the PACF displays a sharp cutoff while the ACF decays slowly, it indicates the presence of AR terms. Similarly, if the ACF displays a sharp cut-off while the PACF decays slowly then it indicates presence of MA terms. If there is no clarity in finding order of ARIMA, more than one model is checked to see which ARIMA model gives the lowest Akaike Information Criterion (AIC). AIC is an estimator of the relative quality of statistical models for a given set of data.

3. Diagnostic check: Once, the ARIMA model is identified, further tests are conducted on the residuals of the estimated model to see if they are white noise. To begin with, residuals are checked with ACF and PACF plots to find out whether there exists significant lags or not. No significant lags show that the residuals are white noise. For further confirmation, Ljung-Box Test is used that checks for autocorrelation between lags of residuals. Here, null hypothesis is written as “there is no autocorrelation in the lags of residuals”. If p-values are above 0.05, it means that there is no evidence to reject the null hypothesis. In other words, residuals are white noise, thus confirming the adequacy of the estimated ARIMA model.

4. Forecasting: Once the ARIMA model is identified, it can be used further for forecasting the future values of the time series.

3.2.2 Econometric technique: Generalised AutoRegressive Conditional Heteroscedasticity (GARCH)

ARIMA models discussed in the previous subsection are used to model the conditional expectation of a process given the past data and past errors. It is a method to model the data linearly when variance remains constant. Although, it provides the best linear forecast for the series, it is incapable of handling non-linearity as in the case of stock market volatility. For instance, if it is observed that in the recent past, the returns show high volatility, it is expected that the volatility will be high in the coming days. However, time series models like ARIMA will not be able to capture this because these models assume that the conditional variance is constant. So, time series models that can capture non-linearity should be used.

Before the introduction of the Auto Regressive Conditional Heteroskedasticity (ARCH) model by Robert Engle in 1982, econometricians described variances of models by using rolling standard deviation by giving equal weights to standard deviation corresponding to past observations. Mathematically this equal weighted model can be written as:

$$\sigma_{u_t} = \frac{1}{N}\sum_{t=1}^{N} \sigma_{u_{t-1}} \tag{3.5}$$

However issue with this approach is that past innovations are given equal weightages which conflicts with the fact that more recent observations are more likely to be more relevant. Also, restricting weights only to some arbitrary number of observations is also not ideal. To get rid of this issues, ARCH models assume that the variance of the current error term or innovation to be a function of the previous time periods' error terms and instead of weighting each value equally, ARCH treats the weights as parameters to be estimated. ARCH allows the conditional variance to change over time as a function of past errors leaving the unconditional variance constant. With this, the ARCH(m) model of order m can be written as:

$$u_t^2 = c + \sum_{i=1}^{m} \alpha_i u_{t-i}^2 + w_t \tag{3.6}$$

where w_t is some white noise process. Although, ARCH model is a better model compared to the model with equal weights to past errors, it too has some limitations. ARCH model assumes that positive and negative shocks have the same effects on volatility as it depends on the square of the previous shocks. In practice, it is well known that price of a financial asset responds differently to negative and positive shocks. It is also known to be a restrictive model especially with respect to constraints for higher order ARCH models. They have very short persistence and respond slowly to large isolated shocks to the return series. To overcome these limitations in the ARCH model, Bollerslev, 1986 introduced Generalized Autoregressive Conditional Heteroscedastic (GARCH) model that allowed much more flexibility in lag structure compared to ARCH. GARCH is a time series modelling technique that uses past variances and past variance forecasts to forecast future variances. Conditional here implies dependence on the observations of the immediate past, and autoregressive describes a feedback mechanism that incorporates past observations into the present. Let ε_t denote a real-valued discrete-time stochastic process and ψ_t denote the information set of all information through time t, then the generalised GARCH(p,q) model with p being the number of autoregressive lags (ARCH terms) and q being the number of moving average lags (GARCH terms), ε^2 can be written as:

$$\varepsilon_t | \psi_{t-1} \sim N(0, h_t), \tag{3.7}$$

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}^2 \tag{3.8}$$

where,

$$p \geq 0, \quad q > 0$$
$$\alpha_0 > 0, \quad \alpha_i \geq 0 \quad i = 1, 2...., q$$
$$\beta_i \geq 0, \quad i = 1, 2...., q$$

GARCH(1,1) is the simplest and one of the most widely used volatility models. Here, with p=1 and q=1, GARCH(1,1) model can be written as:

$$h_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}^2 \tag{3.9}$$

where, $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 > 0$ and $\alpha_1 + \beta_1 < 1$

To determine the ARCH and GARCH order terms, GARCH models can be treated as ARMA/ ARIMA model for squared residuals and then traditional model selection criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) can be used. Model with the lowest AIC and BIC value is selected. There are many extensions of the generalized GARCH model used in volatility modelling depending on the characteristic of the volatility to be modeled. Some of these include Exponential GARCH (EGARCH), Threshold GARCH (TARCH) , Quadratic GARCH (QGARCH), etc.

GARCH Procedure for implementation: Since GARCH uses ARIMA model in its specification, one starts with the ARIMA modelling to model returns and then examines squared residuals obtained from ARIMA model to see whether there are ARCH effects. If arch effects are found in the squared residuals, GARCH modelling can be applied. Although there are variety of GARCH models available, GARCH(1,1) model is used widely because of its simplicity as well as effectiveness. One may check multiple models and finally determine the order of p and q in GARCH using AIC and BIC criterion. Diagnostic plots like ACF, PACF, QQ-plot (for normality check), etc. on squared residuals are used to check appropriateness of identified GARCH model. Also, Ljung-Box Test and ARCH LM tests are performed to test presence or absence of serial correlation in the squared residuals obtained from GARCH model thereby validating GARCH model adequacy.

3.2.3 Statistical technique: Regression analysis

Regression analysis is a statistical technique used to estimate whether a certain variable or variables called as independent variables will have an impact on another variable called dependent variable. Regression analysis produces a regression equation wherein coefficients represent the relationship between each independent variable and the dependent variable. This equation can be used for making predictions. There are different types of regression analysis techniques like Linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression, etc. Since, linear regression is used in this thesis it is discussed here in more detail:

In a linear regression model, the variable of interest called the dependent variable is predicted from k other variables called independent variables using a linear equation. If Y_t denotes the dependent variable, and $X_{i's}$ $(X_{1t}, \ , X_{kt})$ are the independent variables, the value of Y_t can be determined by the linear equation:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + \varepsilon_t \tag{3.10}$$

where the betas (β) are constants and the epsilon is independent and identically distributed (i.i.d.) normal random variable with mean zero called as regression residual. β_0 represents the intercept of the model which is the expected value of Y_t when all the X_{is} are zero. β_{is} are the coefficient (multiplier) of the variables X_{its}.

For interpreting results of regression output, the researcher has to focus on the following statistical measures:

- R-squared or Adjusted R-squared (in case of multiple regression): R-squared measures the proportion of the variation in the dependent variable (Y) explained by the independent variables (X) in the linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model and is more appropriate to considered when there are multiple variables. If R-squared is 0.7, it means 70% of the variation in the output variable is explained by the input variable. However, one should also look at the the problem being considered for which a lower R-squared can still be considered appropropriate. In short, R-squared provides a measure of strength of relationship between dependent variable and independent variables; however it does not comment on whether the relationship is statistically significant or not.

- F-test: F-test in regression compares compares a model with no predictors (independent variables) to the model that is specified by the researcher. A regression model without predictors (independent variables) is also known as an intercept-only model. Unlike t-tests that assesses only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously and then finds whether addition of these variables together in the model is significant enough for them to be there or not. This is done by finding whether p-value (probability) of F-statistic is significant or not for the following hypothesis:

  H_0 : The fit of intercept only model and the specified model is same. i.e. additional variables taken together do not make model significantly better

H_a : The fit of intercept only model is significantly less compared to the specified model. i.e. additional variables taken together do make the model significantly better.

If p-value of F-statistic is extremely small (less than 0.05), we can reject H_0 at 5% significant level and say that overall addition of variables is significantly improving the model.

- Standardized coefficient (β): Standardized regression coefficient (β) is one of the most frequently reported summary statistics used with multiple regression. β represents an amount of change that occurs in the predicted value of the dependent variable as a result of a change in an independent variable, assuming all other independent variables remain constant, with the changes expressed in standardized form.

  β compares the strength of the effect of each individual independent variable to the dependent variable and higher the absolute value of the beta coefficient, the stronger the effect. For example, a β of -.8 has a stronger effect than a β of +.7. βs are measured in standard deviations as their units and hence can be compared to one another easily. If β of the independent variable X is positive, it indicates that for every unit increase in the independent variable, the dependent variable will increase by the value of β. However, if β of the independent variable X is negative, it indicates that for every unit increase in the independent variable, the dependent variable will decrease by the value of β.

- Standard error: The standard error in the regression output is a measure of the precision of the model. It represents the average distance that the observed values fall from the regression line. In other words, it indicates the average error of the regression model implying how wrong regression model could be to make predictions and hence for a regression model to be relied on, standard error should be very low. If approximately 95 of the observations fall within $+/-2$*standard error of the regression from the regression line, it can be considered as a quick approximation of a 95% prediction interval.

- t-statistic or the t-value: It is computed by dividing the coefficient by its standard error. A larger coefficient value compared to standard error makes the t-value larger thereby indicating a more reliable coefficient. t value by itself may not be conclusive but is needed to compute the p-value.

- p-value: The p-value for each predictors in the regression tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value ($<$0.05) indicates that one can reject the null hypothesis suggesting that the predictor is a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. On the other hand, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response. The t value is used to look up the Student's t distribution to determine the p-value.

Linear regression makes several assumptions about the data as given below :

- Linearity of the data: The relationship between the predictor (X) and the outcome (Y) is assumed to be linear.
- Normality of residuals: The residual errors are assumed to be normally distributed.
- Homogeneity of residuals variance: The residuals are assumed to have a constant variance (homoscedasticity)
- Independence of residuals error terms: Residual errors are assumed to have no autocorrelation with themselves.

The assumptions of linear regression can be tested by using regression diagnostic plots. In this thesis, the diagnostic plots are used to test assumptions of regression model consisting of the dependent variable Returns (Nifty50) and the independent variable Sentiment (aggregate news sentiment). The diagnostic plot along with its interpretation for each of the assumptions is given below:

- Test of linearity of data: This test is conducted by plotting Residuals vs Fitted values. Figure 3.2 shows no distinctive pattern and major portion of the solid line (in red) through the scattered plot is approximately straight and horizontal. This suggests that relationship between the predictors and the outcome variables is fairly linear. .

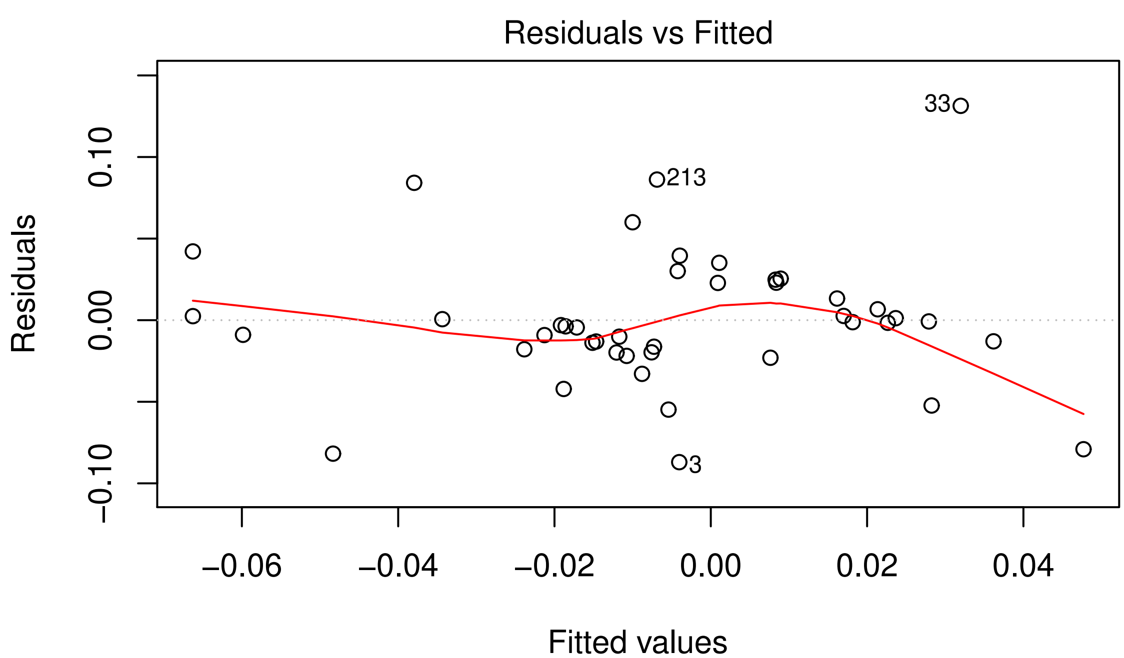


Figure 3.2: Test of Linearity: Residuals vs Fitted

- Test of normality: The normality assumption is evaluated using QQ-plot by comparing the residuals to ideal normal observations along the 45-degree line. QQ-plot shown in Figure 3.16 indicates that the residuals follow approximately a straight line along 45-degree line. This suggests that standardized residuals have normal distribution. .

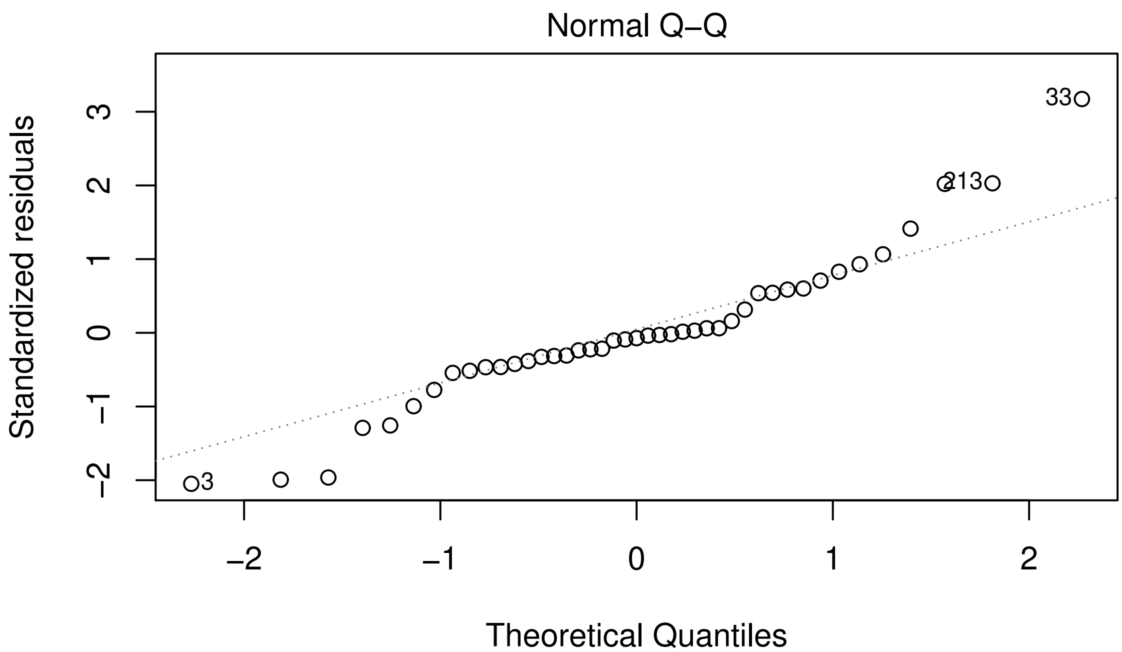


Figure 3.3: Test of Normality: Normal Q-Q plot

- Test for homogeneity of residual variance: The assumption of homogeneity of residual variance can be checked by examining the scale-location plot, also known as the spread-location plot. This plot should show residuals spread equally along the ranges of predictors for confirming homogeneity of residual variance. As shown in Figure 3.4, standardized residuals are approximately spread equally along the range of predictors and the red line is approximately horizontal and flat thereby indicating homogeneity of residual variance.

- Test of independence of residuals error terms: To test the independence of residual errors, Durbin-Watson test for autocorrelation of disturbances is used. This test uses the following hypothesis:

  H_0 (null hypothesis): There is no correlation among the residual errors.

  H_a (alternative hypothesis): The residual errors are autocorrelated.

  The result of this test is given below:

```
lag Autocorrelation D-W Statistic p-value
```

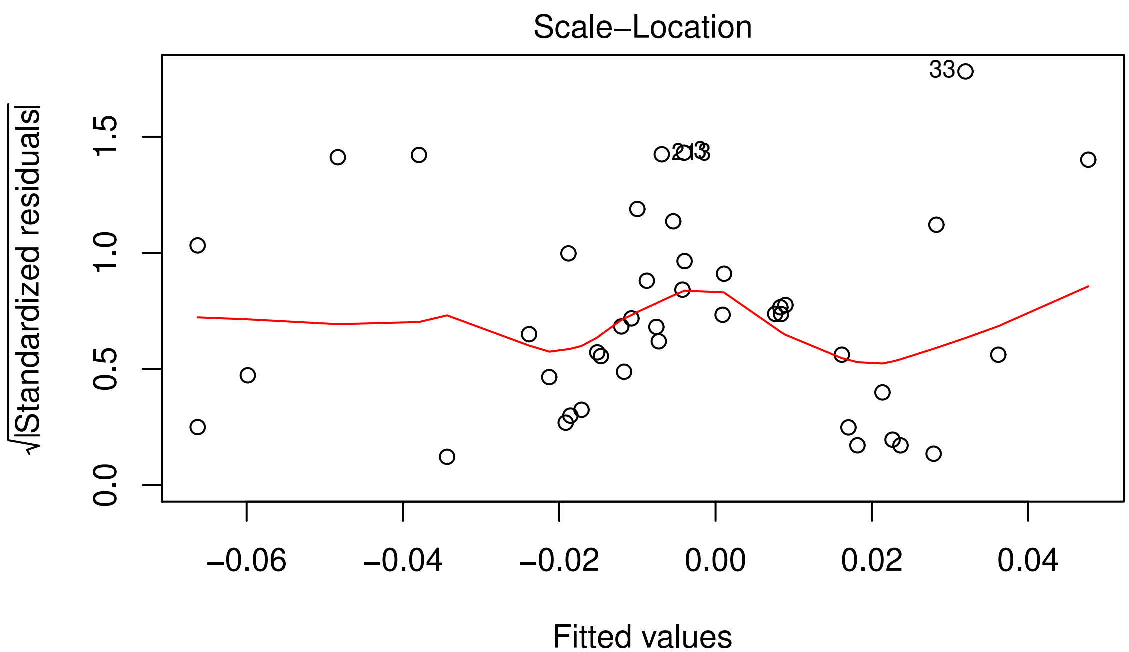


Figure 3.4: Test of Homogeneity: Scale-Location Plot

```
1     -0.02069272      1.842488   0.534
Alternative hypothesis: rho != 0
```

The result shows that p-value is greater than 0.05 indicating that there is no evidence to reject the null hypothesis. In other words, the residual errors are not autocorrelated and hence residual errors are independent.

3.2.4 Text mining technique: Sentiment Analysis

Text mining is a process of examining large collections of text to discover new information or help answer specific research questions. Text mining has gained enormous importance today because it can process massive amount of unstructured text data in the form of emails, social media posts, news articles, etc. The two most common methods used in text mining include text classification and text extraction. Text classification uses techniques like topic analysis, sentiment analysis, language detection, and intent detection. On the other hand, text extraction uses techniques like keyword extraction, named entity recognition, and feature extraction.

In this thesis, sentiment analysis is used based on the dictionary approach. Some of the important characteristics of this implementation[1] are stated below:

- Dictionary used: Henry's finance dictionary (HE)

[1]Details of implementation are given in subsections 3.3.5 and 3.3.6

- Text: News articles from economictimes.indiatimes.com portal
- Pre-processing of text: This includes steps for converting html text to plain text, removing stop words, extra spaces, and punctuation marks, stemming, converting text to lower case.
- Features: Semantic features - Positive and negative sentiment words, syntactic feature - Bag of words (bigrams), Intensifiers - amplify and deamplify
- Sentiment polarity formula:

$$S_d = \sum_{i=1}^{Q_d} w_i v_i s_i \tag{3.11}$$

where S_d is the sentiment of a news article d, Q_d is the number of unigrams in document d, v_i is the impact of a valence shifter corresponding to shifting value of the preceding unigram $(i-1)$.

Validity of Sentiment analysis using Henry's finance dictionary

In this thesis, sentiment analysis based on Henry's finance dictionary is used to derive sentiment of each news article. While this section discusses the validity of using Henry's dictionary in sentiment analysis, the Section 3.3 provides the details of how sentiment analysis is applied to derive sentiment of news articles. Before validating the use of Henry's dictionary in sentiment analysis, it also important to validate the use of dictionary approach compared to other approaches. Literature review conducted in the previous chapters reveals two broad approaches being used in sentiment analysis viz. Machine learning approach and Lexicon based approach. While other approaches can also be used, lexicon based approach using pre-specified dictionary has some advantages as given below:

- Unlike supervised machine learning approach, no training data is required.
- Since human annotators are not involved in classification process, research subjectivity is avoided when classifying the sentiment of different words.
- Scalable over volume of text data.
- Since dictionaries are publicly available, research can be easily replicated by others.
- Dictionary based approach is parsimonious as compared to other sentiment analysis approaches which use manually trained data and complex algorithms.

Choice of Henry's finance dictionary: The choice of a dictionary is an important decision a researcher has to take in order to conduct research using sentiment analysis. In the past, researchers have used various dictionaries suitable for their study. Some popular dictionaries as cited in Loughran and Mcdonald (2016) used in sentiment analysis related to accounting and finance are compared below:

- Harvard General Inquirer dictionary: This is a general purpose dictionary developed by Harvard University with a list of positive and negative words and is also used in the General Inquirer software. Along with diction word list, it is one of the earliest dictionaries used for sentiment analysis across various domains including finance. It consists of 2005 negative words and 1637 positive words. One of the disadvantages of using this dictionary in finance domain is that it is a general purpose dictionary and may not identify words from financial text accurately. Loughran and Mcdonald (2011) finds that almost three-fourths (73.8%) of the negative word counts in this dictionary are words that are typically not negative in a financial context.

- Henry's Finance dictionary: This is one of the earliest dictionaries created specifically for financial texts and was created by examining earnings press releases in the telecommunication and computer-service industries (Loughran & Mcdonald, 2016). It contains 105 positive and 85 negative words which is one of the limitations of this dictionary. List of positive and negative words from Henry's finance dictionary is given in Appendix A.

- Loughran-McDonald financial sentiment dictionary: This is one of the most extensive dictionaries created specifically for sentiment analysis of accounting and finance text. It is created by using 10-Ks[2] and hence probability of identifying words from business context is high. It consists of 2355 negative words and 354 positive words. While this dictionary is very extensive, the proportion of negative words compared to positive words is large which may introduce a bias towards negative words as the probability of matching negative word from text is higher compared to the probability of matching positive words because of availability of larger set of negative words in the dictionary.

- Diction Optimism and Pessimism Word List: Along with the Harvard General Inquirer dictionary, this is one of the earliest dictionaries used for sentiment analysis. According to Loughran and Mcdonald (2016), this dictionary has 35 subcategories such as praise, satisfaction, inspiration, blame, hardship, denial, etc. To get a list of positive words, subcategories praise, satisfaction, and inspiration are combined to create 686 Diction optimism words (positive words). Similarly, words from the subcategories blame, hardship, and denial are combined to create 920 Diction pessimism words (negative words). Similar to Harvard General Inquirer dictionary it is a general purpose dictionary and hence may not

[2] A 10-K is a comprehensive report filed annually by public companies about their financial performance and is required by the U.S. Securities and Exchange Commission (SEC)

give accurate results when used specifically for sentiment analysis of text in the finance domain.

Henry and Leone (2011) suggested use of domain specific wordlist in sentiment analysis of qualitative content of financial disclosure can significantly increase the power for their tests. A comparison of the four dictionaries compared above reveal that only Henry's Finance dictionary and Loughran-McDonald financial sentiment dictionary are the most suitable dictionaries for sentiment analysis of text in finance domain. While Henry's Sentiment dictionary has a limited set of words, Loughran-McDonald financial dictionary suffers from bias towards negative words as explained earlier. Moreover, Meier et al. (2018) reported that for sentiment analysis, it does not seem to matter whether the extensive dictionary by Loughran and McDonald or the much shorter dictionary by Henry is used; the quality measures of both dictionaries seem to differ only marginally. So, even though Henry's dictionary has a less number of positive and negative words, comparatively it is more suitable for studies that examines text for both negative and positive sentiment. Kearney and Liu (2014) reported the use of Henry's finance dictionary (HE) in Henry(2008), Henry and Leone (2009), Rogers et al. (2011), Price et al. (2012), etc.

3.3 Method

Most of the studies on news sentiment in the past are conducted by following a methodology wherein the relationship between news sentiment and stock variables like returns, volatility, trading volume, etc. is examined by collecting stock's data corresponding to the time when news related to its company is released. While the noise trader theory indicates presence of noise traders during the period when prices deviate from their intrinsic value due to sentiments and limits on arbitrage, studies on news sentiment do not explicitly consider data points based on extreme deviations caused by the news flow. So, these studies do not control for the movement due to fundamentals of the stock. For instance, these studies collect news on events like merger and acquisition, earnings announcement, etc. and then examine their impact on the stock variables irrespective of whether there was any extreme deviation in prices caused when the news was released.

According to noise trader theory, limits to arbitrage and sentiments are the reasons for stock prices to deviate away from intrinsic value. This thesis is based on the premise that extreme deviation in market price is caused by investor sentiment as described in the noise trader theory and examines the role of aggregate news sentiment in triggering investors' sentiment to drive market to extreme levels. The focus is on identifying trading days with extreme returns and examining them along with days around them to see whether news sentiment obtained from news published related to business, economy, politics and international categories have any role in influencing stock market returns. Steps used in this methodology are given in the following subsections:

3.3.1 Examination of stationarity and description of data

Nifty50 index price data is a time series data and if any model has to be applied on it, one should ensure that it is stationary. A time series is said to be stationary if its statistical properties like mean, variance and auto-correlation structure do not change over time. So, Nifty50 index price series is checked to see if it is stationary. This examination is done by using Augmented Dickey-Fuller (ADF) Test. Result of this test is given below:

```
Augmented Dickey-Fuller Test
data:  nif50
Dickey-Fuller = -3.2558, Lag order = 14, p-value = 0.07826
alternative hypothesis: stationary
```

Since the p-value (0.07826) is not significant at 5% level, we fail to reject the null hypothesis that the Nifty50 index price series is not stationary. Hence in order to make the series stationary, difference of its logarithmic price lags is obtained to give logarithmic returns as shown in the following equation 3.12:

$$R_t = \log(P_t) - \log(P_{t-1}) \tag{3.12}$$

where R_t is daily logarithmic return at time t, P_{t-1} and P_t are Nifty50 daily adjusted closing prices on two successive trading days $(t-1)$ and (t) respectively. The differenced series of the log prices also represents continuously compounded returns or log returns (hereafter referred as returns). Plot of Nifty50 index returns is shown in Figure 3.5. The return series is again checked for stationarity using ADF test. Result of ADF test on the Nifty50 index return series is given below:

```
Augmented Dickey-Fuller Test
data:  resret
Dickey-Fuller = -13.751, Lag order = 14, p-value = 0.01
alternative hypothesis: stationary
```

The p-value (0.01) here is significant at 5% level. So, the null hypothesis that the returns series is not stationary is rejected. In other words, the Nifty50 index return series is stationary and can be used for further modeling.

Descriptive statistics

Apart from examination of stationarity in a time series, before conducting any analysis on data, it is appropriate to understand some of the important statistical properties of this data. Table 3.1 presents descriptive statistics of Nifty50 index returns.

Measure of central tendency: The measures of central tendency viz. mean and median have a value of 0 as shown in the Table 3.1 indicating that the distribution of Nifty50 returns is symmetric.

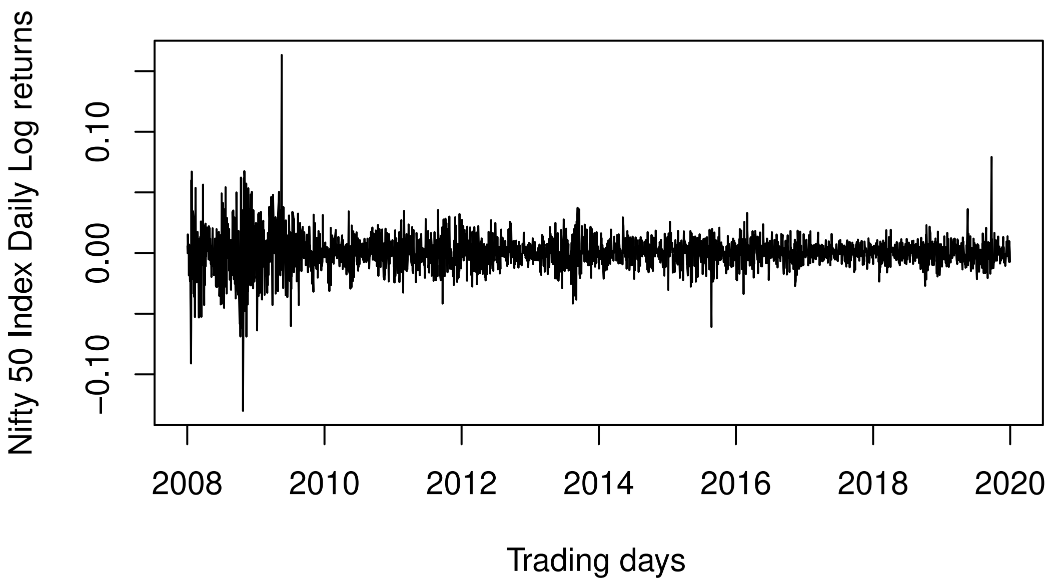


Figure 3.5: Nifty50 Index daily log returns

Table 3.1: Summary Statistics of Nifty50 returns

Statistic	Nifty50 returns
n	2955.000
mean	0.000
sd	0.014
median	0.000
min	-0.130
max	0.163
range	0.293
skew	0.141
kurtosis	13.922
se	0.000

Measure of dispersion: The measure of dispersion viz. standard deviation (0.014) and standard error (0) shown in Table 3.1 have values close to zero indicating that the data points are close to the mean without much spread in the data. Low spread in data with large number of data points being close to zero can also be understood from the range of 0.293 with maximum of 0.163 and minimum of -0.130; however presence of some outliers or extreme values can mislead this interpretation. The presence of extreme values can be roughly understood from the kurtosis which is a measure of peakedness in data distribution. As compared to the kurtosis of 3 for normal distribution, this data has a kurtosis of 13.92196 which shows peakedness in the data thereby indicating presence of some extreme data points. Skewness of 0.141 is close to zero indicating a symmetrical distribution of data but slightly skewed to the left.

3.3.2 Filtering trading dates with extreme returns

The basic premise this study is based on is that the noise traders are highly active when stock market prices start moving away from intrinsic value to extremes and are influenced by the sentiments generated as a result of the news that they receive on the events happening in the global and domestic market environment. In order to examine this idea, a sample consisting of data points with extreme deviations need to be obtained. In order to get an appropriate sample, Nifty50 returns obtained over a 12 year period for 2995 trading days is examined by plotting it in a QQ-plot to check whether it follows a normal distribution.

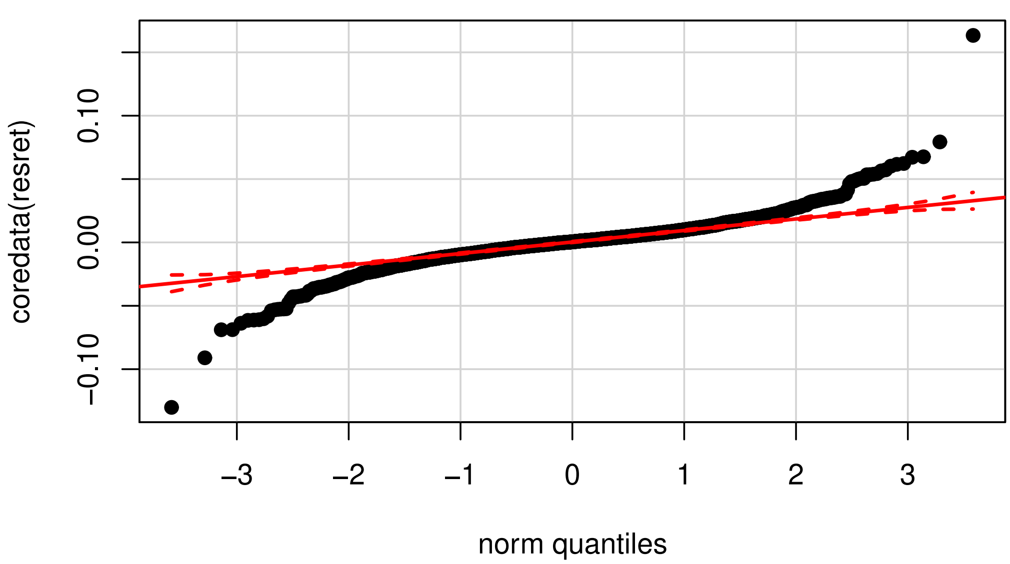


Figure 3.6: QQ Plot of Nifty50 returns

Visual inspection of the plot in Figure 3.6 indicates that the data is approximately normally distributed as most of the points lie along the 45 degree line. However, it also shows a possibility of having some data points with extreme values.

In this thesis, 1% extreme cases are considered by setting a confidence interval of 99%. 1% cases out of 2995 corresponds to 29.95 which is approximately 30 data points. In statistics, for a normal distribution, a minimum sample size of 30 is considered to be adequate. The Nifty50 index returns distribution being approximately normal with the expectation that there are going to be extra data points in the tail as indicated by QQ-plot, the sample size expected is more than 30. In a normal distribution, a confidence interval of 99% corresponds to a z-value of 2.58. So, a cut-off at (+/-) 2.58σ is set around the returns to examine volatility. So, the extreme returns are defined as Nifty50 index log returns beyond a cut-off set at (+/-2.58)σ on these returns.

Although, the method of using a cut-off at (+/-) 2.58σ seems to serve the purpose for deriving the sample, using a sample standard deviation σ which is a point estimate is not appropriate as the return series shown in Figure 3.5 shows volatility clustering i.e. there is a presence of sustained periods of high or low volatility. In other words, this return series is heteroscedastic. Also, according to Engle (2001), conventional methods using least square regression to model volatility do not seem to work optimally because they assume that the expected value of all error terms, when squared, is the same at any given point. So, a method that estimates daily σ is a better way to deal with this problem.

ARCH (Auto Regressive Conditional Heteroscedasticity) model and GARCH (Generalised Autoregressive Conditional Heteroscedasticity) models treat heteroscedasticity as a variance to be modeled and hence prediction is computed for the variance of each error term whereby the deficiencies of least squares are corrected. While ARCH uses past variance forecast for modeling future variances, GARCH uses past variances as well as past variance forecasts to forecast future variances. Conditional here implies dependence on the observations of the immediate past, and autoregressive describes a feedback mechanism that incorporates past observations into the present.

In ARCH model, conditional variance σ_t^2 is expressed as:

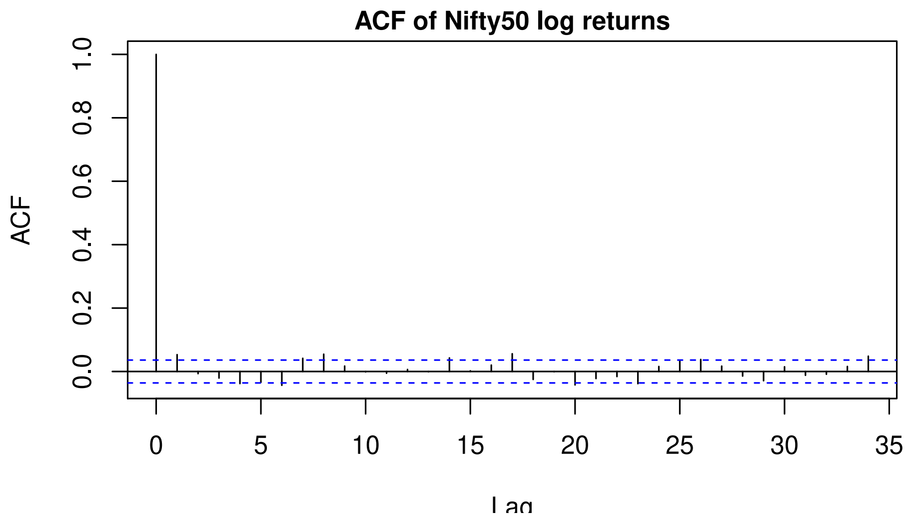
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \tag{3.13}$$

Similarly, GARCH(p,q) model, conditional variance h_t^2 is written as:

$$h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}^2 \tag{3.14}$$

where $\alpha_0 > 0$, $\alpha_i > 0, i = 1, ..., q$; $\beta_j > 0, j = 1, ..., p$; and $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$ for ensuring $\{\sigma_t^2\}$ as weak stationary. Since, Nifty50 index returns series is stationary, it can be modeled using GARCH. However, before modeling volatility, the returns series must be modeled. This can be done using ARIMA. Preliminary inspection of Autocorrelation (ACF) and Partial autocorrelation

(PACF) plots of Nifty50 index returns in Figure 3.7 and Figure 3.8 reveals that ACF cuts off at 1, 7, 8 and 14 and PACF tails off indicating possibility of MA(1) process (considering parsimony of the model by rejecting higher lags at 7, 8 and 14).



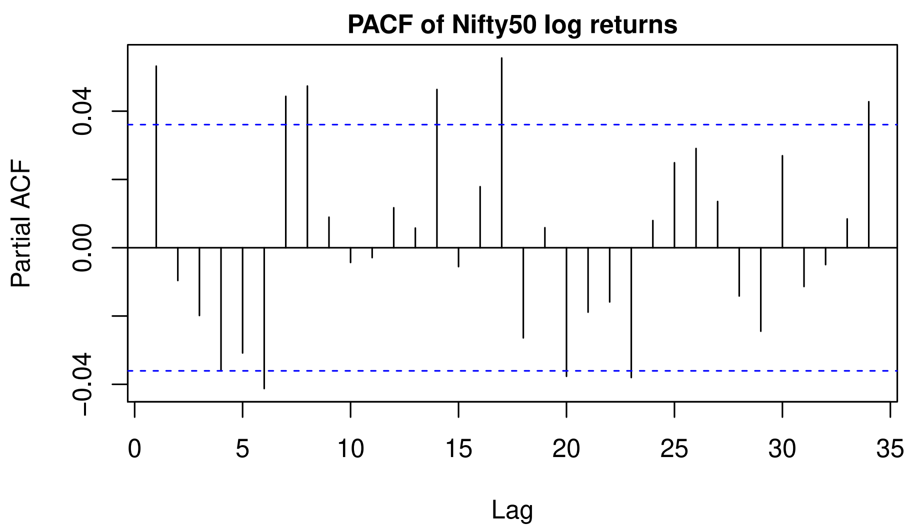


Figure 3.8: PACF plot of Nifty50 index returns

To get clarity on the appropriate order of p, d and q, different models using various combinations of p and q are examined as potential ARIMA models and then compared as shown in Table 3.2. ARIMA(0,0,1) with zero mean is found be the best model based on lowest value of AIC (AIC= -17040.83). This is followed by a test for validating ARIMA model by checking presence or absence of serial correlation in the residuals obtained from ARIMA model. Here, visual examination of ACF and PACF plots on ARIMA(0,0,1) model residuals along with Ljung-Box Test are performed. Plots of ACF and PACF on ARIMA(0,0,1) model residuals are

Table 3.2: Comparison of different orders of ARIMA model

ARIMA model	AIC
ARIMA(2,0,2) with non-zero mean	-17032.11
ARIMA(0,0,0) with non-zero mean	-17033.11
ARIMA(1,0,0) with non-zero mean	-17038.65
ARIMA(0,0,1) with non-zero mean	-17039.61
ARIMA(0,0,0) with zero mean	-17034.25
ARIMA(1,0,1) with non-zero mean	-17036.83
ARIMA(0,0,2) with non-zero mean	-17037.68
ARIMA(1,0,2) with non-zero mean	-17034.87
ARIMA(0,0,1) with zero mean	-17040.83
ARIMA(1,0,1) with zero mean	-17038.02
ARIMA(0,0,2) with zero mean	-17038.9
ARIMA(1,0,2) with zero mean	-17036.1

given in Figure 3.9 and 3.10 respectively followed by Ljung-Box test for confirmation.

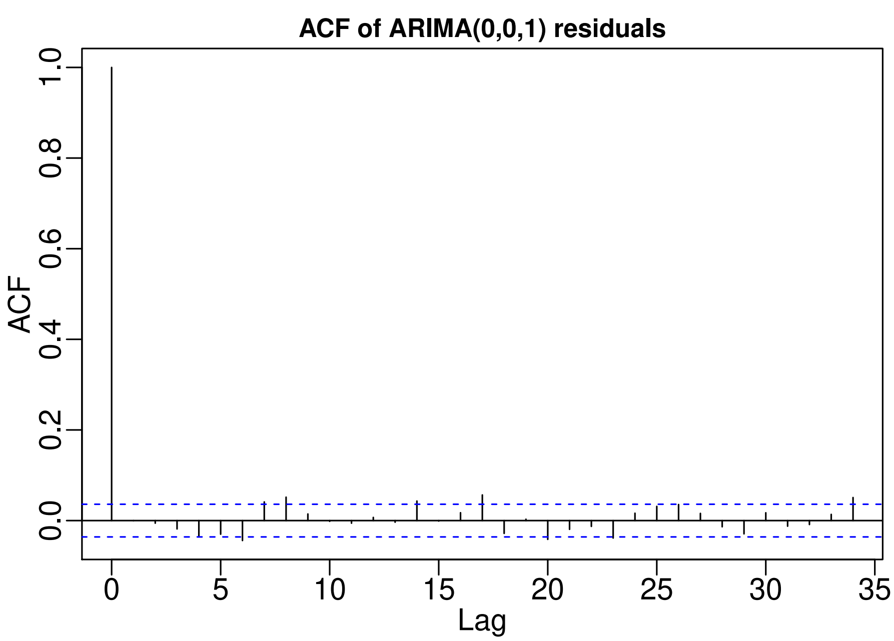


Figure 3.9: ACF: residuals of ARIMA(0,0,1)

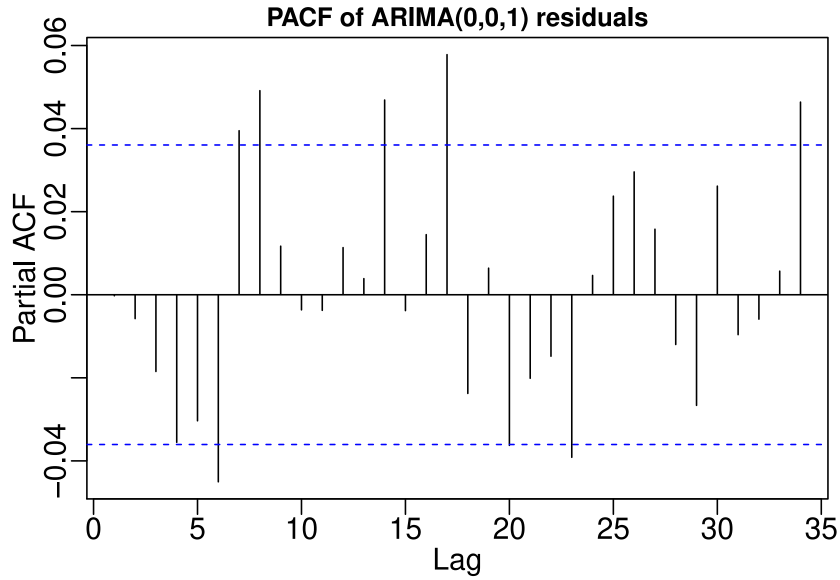


Figure 3.10: PACF: Residuals of ARIMA(0,0,1)

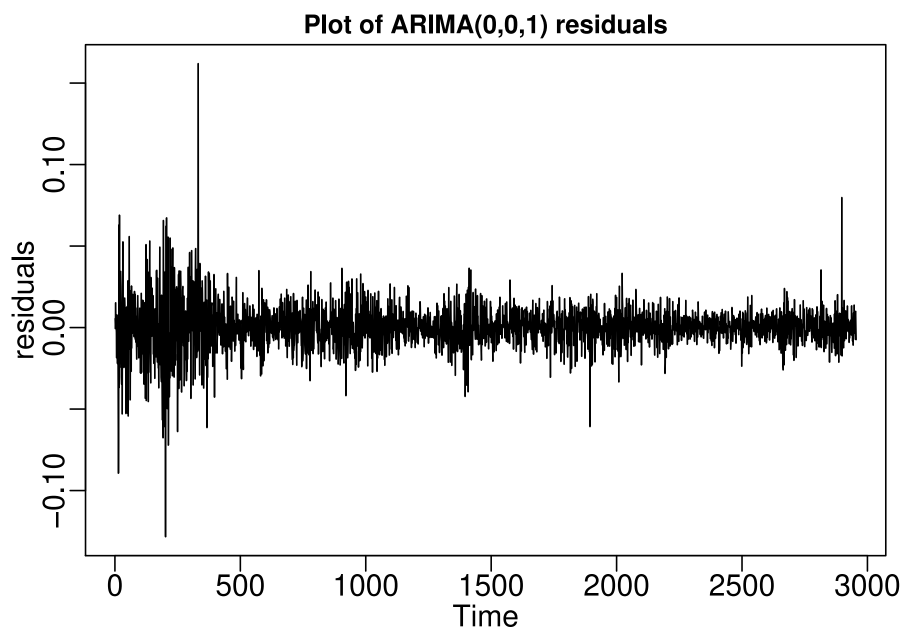


Figure 3.11: ARIMA(0,0,1) residuals

Figure 3.9 and Figure 3.10 show that residuals have no significant lags (considering parsimony of the model by rejecting higher lags), but the time series plot of residuals in Figure 3.11 shows some clusters of volatility. This is because ARIMA is a method to model data linearly and we still need to model volatility. Confirmation of model validity is done using Ljung-Box test as shown below:

```
Box-Ljung test
data:  arima001$residuals
X-squared = 7.5071, df = 5, p-value = 0.1856
```

Result of this test shows no autocorrelations among the model residuals (p-value = 0.1856) upto 5 lags, thus confirming adequacy of the model. ACF seems to die down and PACF cuts down after lag 11 as shown in Figure 3.12 and Figure 3.13 respectively. Also, squared residuals of ARIMA(0,0,1) in Figure 3.14 show volatility clusters.

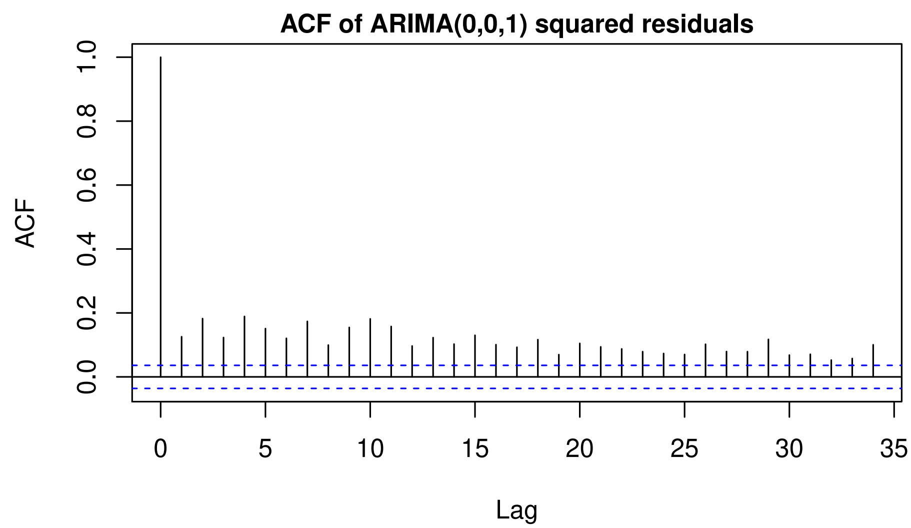


Figure 3.12: ACF: Squared residuals of ARIMA(0,0,1)

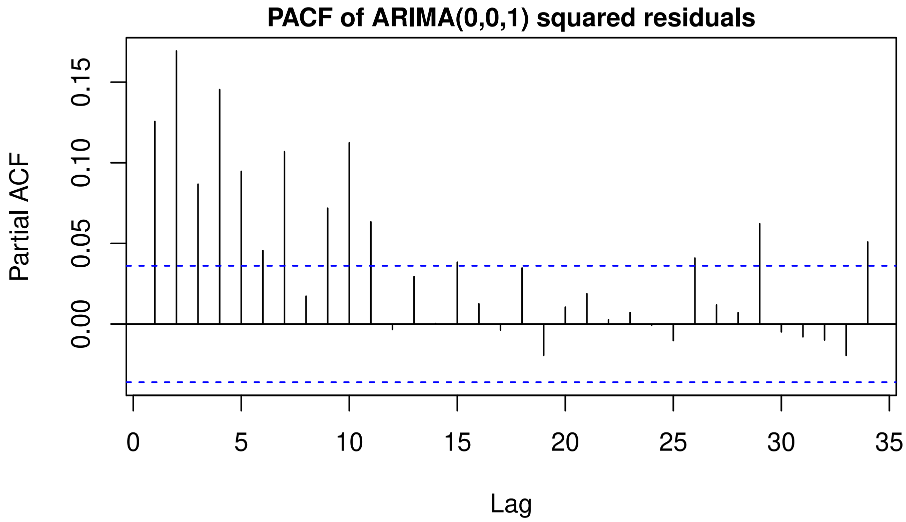


Figure 3.13: PACF: Squared residuals of ARIMA(0,0,1)

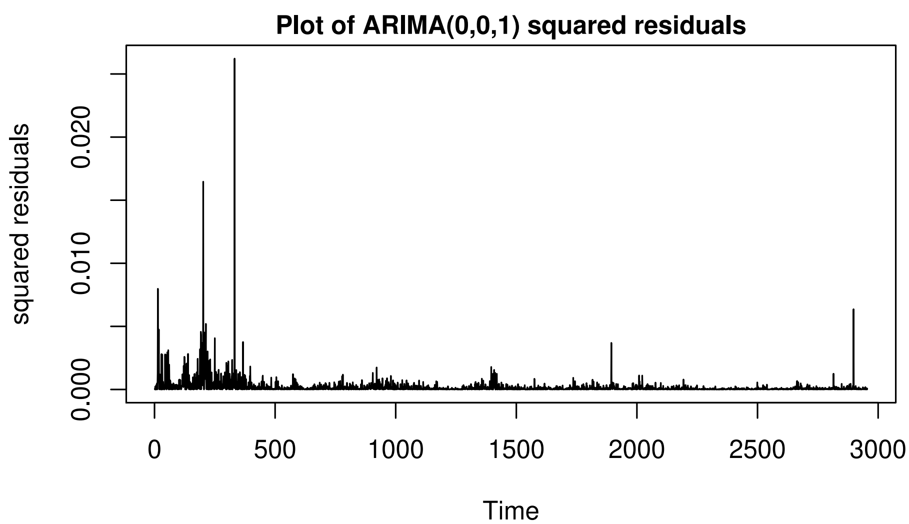


Figure 3.14: ARIMA(0,0,1) squared residuals

Ljung-Box test is also done on squared residuals of ARIMA(0,0,1) to test arch effects in the squared residuals as shown below:

```
Box-Ljung test
data:  residuals(arima001)^2
X-squared = 363.68, df = 5, p-value < 2.2e-16
```

Ljung-Box test on squared residuals of ARIMA(0,0,1) suffers from arch effect as null hypothesis that there is no arch effect is rejected (p-value= 2.2e-16). Presence of arch effects in the squared residuals of ARIMA(0,0,1) indicate autocorrelation among lags of past variance and innovations which can be modeled using GARCH.

Although, GARCH(p,q) model with different values of p and q and other GARCH variants may be examined, GARCH(1,1) model has been used here as it is known to be parsimonious, popular and highly effective as compared to others in modeling volatility. GARCH(1,1) model requires ARIMA model parameters in its specification along with the type of distribution. Since, ARIMA(0,0,1) is found to be appropriate to model Nifty50 Index returns, p and q parameters are used in GARCH specification along with t-distribution. To check, appropriateness of GARCH(1,1) model, diagnostic analysis based on Weighted Ljung-Box test and LM test are performed.

Table 3.3a, Table 3.3b and Table 3.3c show that all lags of the model in Weighted Ljung-Box Test on Standardized Residuals, Standardized Squared Residuals and ARCH-LM Test have p-value above 0.05 indicating that there are no arch effect. This means that GARCH(1,1) is appropriate to model volatility in Nifty50 index returns. Validity of GARCH(1,1) model is further confirmed by using diagnostic plots given in Figure 3.15 and Figure 3.16.

Table 3.3: GARCH(1,1) model diagnostic tests

Weighted Ljung-Box Test on Standardized Residuals		
	statistic	p-value
Lag[1]	1.021	0.3122
Lag[2*(p+q)+(p+q)-1][2]	1.076	0.6936
Lag[4*(p+q)+(p+q)-1][5]	1.585	0.8254
d.o.f=1		
H0 : No serial correlation		

(a) Weighted Ljung-Box Test on Standardized Residuals

Weighted Ljung-Box Test on Standardized Squared Residuals		
	statistic	p-value
Lag[1]	0.001065	0.974
Lag[2*(p+q)+(p+q)-1][5]	0.453761	0.9646
Lag[4*(p+q)+(p+q)-1][9]	0.990577	0.9866
d.o.f=2		

(b) Weighted Ljung-Box Test on Standardized Residuals

Weighted ARCH LM Tests				
	Statistic	Shape	Scale	P-Value
ARCH Lag[3]	0.2068	0.5	2	0.6493
ARCH Lag[5]	0.9455	1.44	1.667	0.7492
ARCH Lag[7]	1.1385	2.315	1.543	0.89

(c) Weighted ARCH LM Tests

Acf plot of squared residuals in Figure 3.15 shows that all lags are within the confidence interval indicating no arch effect. QQ plot in Figure 3.16 shows normal distribution. All these diagnostic tests confirm the validity of GARCH(1,1) model.

GARCH(1,1) model provides standard deviation (sigma) of the return series for each data point (trading date). As mentioned above, trading dates with extreme returns are then filtered by setting a cut-off level of (+/-) 2.58σ as shown in Fig. 3.17. This gives a total of 45 data points. Since the cause of the extreme movement is unknown, news published around the day of extreme returns is examined by keeping an observation window. Referring short term event studies like Kabir et al. (1997), Lee (1997), Reuer and Miller (1997), Tang and Tikoo (1999), Reuer (2001) and Hayward (2002) as cited in (Oler et al., 2008), observation window size of 5 trading days is set which includes day of extreme stock market movement (t_0), two trading days before (t_{-1} and t_{-2})

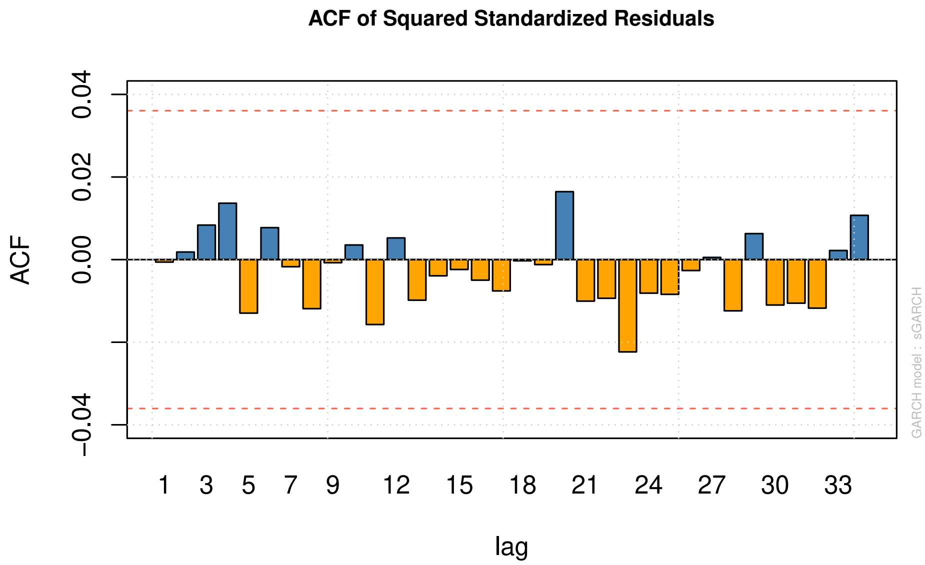


Figure 3.15: ACF of Squared Residuals of GARCH(1,1) model

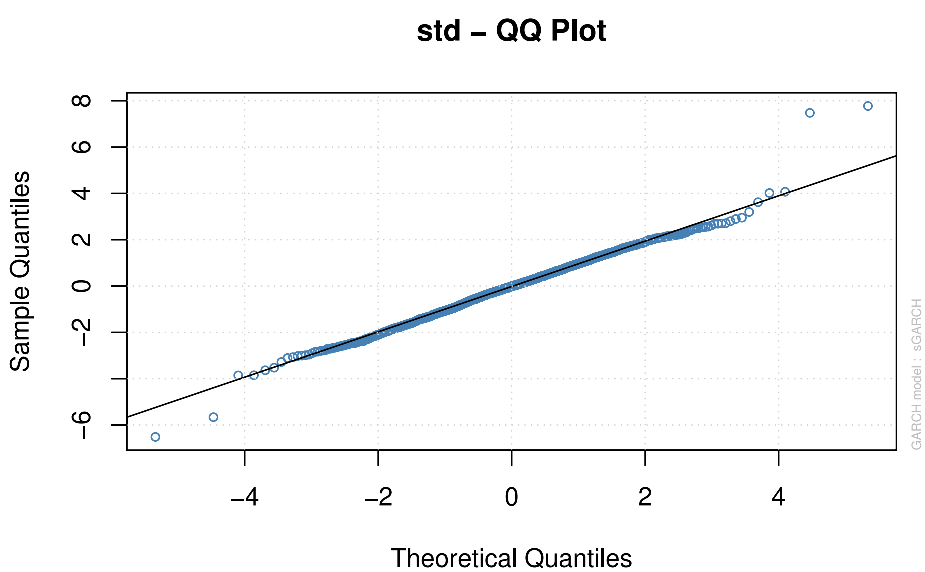


Figure 3.16: QQ-Plot of Standardized Residuals

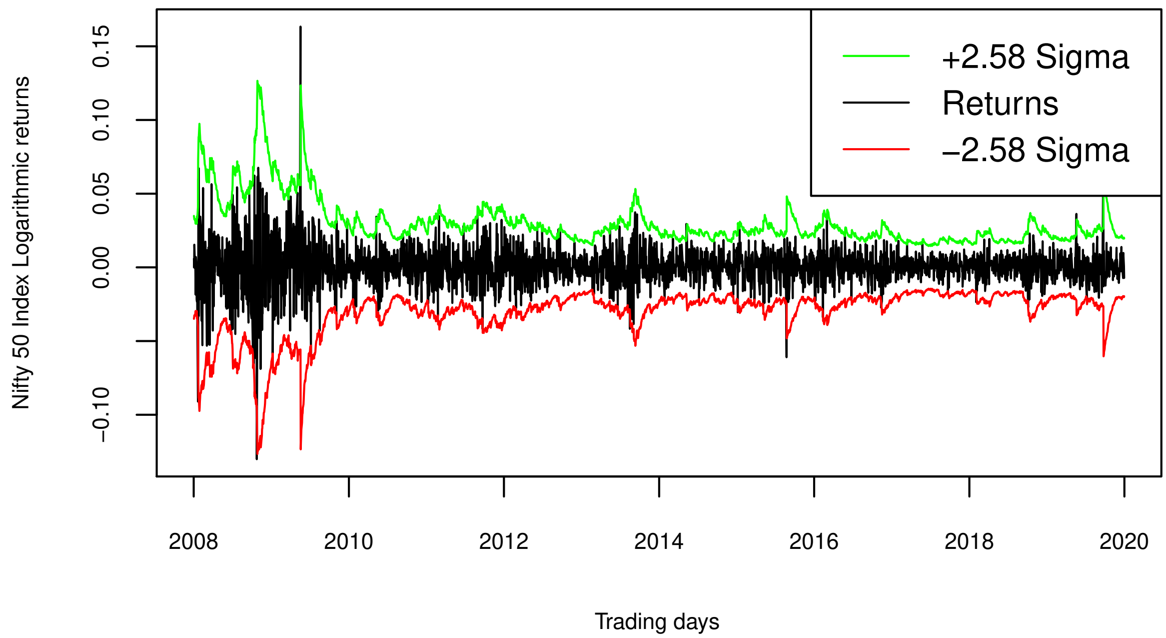


Figure 3.17: Extreme event filtering using conditional volatility

and two trading days after (t_{+1} and t_{+2}) the day of extreme stock market movement. If there is any overlap between observation windows, trading dates obtained as data point that are chronologically older are discarded. Applying this rule, 43 trading dates with extreme returns on Nifty50 index are obtained as shown in Table 3.4. Along with the returns corresponding to these trading dates obtained earlier, returns (logarithmic returns) for other trading days in the observation window corresponding to each of the 43 data point are also obtained.

3.3.3 Obtaining returns of Nifty50 index and Sectoral indices

After filtering trading dates with extreme returns and forming observation windows around them, Nifty50 index returns corresponding to trading days in all observation windows are obtained as a sample for further analysis. Also, daily adjusted closing prices of 10 sectoral indices viz. NIFTY Auto, NIFTY Bank, NIFTY FMCG, NIFTY Energy, NIFTY IT, NIFTY Media, NIFTY Pharma, NIFTY Private Bank, NIFTY PSU Bank and NIFTY Realty corresponding to trading days in all observation windows are obtained from www.niftyindices.com. This price data is further converted into their respective sectoral indices returns using equation 3.12.

Table 3.4: Trading dates with extreme returns

2008-01-21	2008-09-19	2008-10-10	2008-10-24	2009-01-07
2009-03-23	2009-05-18	2009-07-06	2009-11-03	2009-12-23
2010-01-27	2010-05-10	2010-09-13	2010-10-13	2011-01-07
2011-02-24	2011-06-24	2011-08-29	2011-09-22	2012-06-06
2012-06-29	2012-09-07	2012-09-14	2013-02-21	2013-06-20
2013-08-16	2014-05-09	2015-01-06	2015-03-26	2015-05-06
2015-08-24	2016-01-04	2016-02-11	2016-03-01	2016-05-25
2016-06-24	2016-11-11	2017-03-14	2018-02-02	2018-10-04
2019-05-20	2019-07-08	2019-09-24		

3.3.4 Webscraping and news categorization

After obtaining the returns corresponding to trading dates in all observation windows, the next step is to gather news articles. Webscraping process is used to gather news corresponding to those trading dates in all observation windows from the archives of "economictimes.indiatimes.com" portal. However, since stock market is not open on all 7 days of the week and also on public holidays, if there are non-trading days in between trading days in the observation windows, news published on those days are also webscraped. News published on non-trading days are added to the collection of news published on the next trading day. Because of this there are a total of 322 dates for which news articles are webscraped although there are only 215 trading dates corresponding to 43 dates with extreme returns.

Webscraping process starts by obtaining URLs (Uniform Resource Locator) of hyperlinks from the news archive for all dates obtained above. Because of the hierarchical url structure used by the "economictimes.indiatimes.com" portal, news articles are already classified in different categories as can be seen from the examples given below:

- http://economictimes.indiatimes.com/news/politics/nation/work-on-badarpur-flyover-within-next-3-months-hooda/articleshow/2718573.cms
- http://economictimes.indiatimes.com/news/economy/policy/many-companies-still-grappling-with-new-csr-rules-of-companies-act/articleshow/45739293.cms
- http://economictimes.indiatimes.com/news/international/world-news/pakistan-cancels-2015s-first-anti-polio-campaign/articleshow/45761691.cms
- http://economictimes.indiatimes.com/opinion/vedanta/happiness-is-here-and-now/articleshow/3495666.cms

- http://economictimes.indiatimes.com/magazines/corporate-dossier/we-want-our-gods-and-our-leaders-to-observe-and-understand-us/articleshow/3501197.cms

Broader categories present in URL include news, opinions, recommendations, interviews, etc. Only URLs having the word “news” in the URL are filtered from the rest and considered for further analysis. Others which are related to opinions, recommendations, interviews, magazines and articles by column writers on specific topic are filtered out and not considered relevant as they are not tagged as news. URLs related to news further show that they are again classified according to various subjects like economy, industry, politics, company, environment, sports, science, defence, international business, etc. Categories like sports, science, defence, environment, etc. generally relate to articles which focus on general issues but not necessarily related to stock market. So all of them are brought under a general category “others” and excluded from this study. News articles from categories “company”, “industry” and “stocks” are brought into a single category “business” as news from all these categories refer to news articles related to business. A total of 42,485 news articles related to four categories economy, politics, business and international are webscraped and used for further processing. Table 3.5 gives category-wise count of news articles.

Table 3.5: Count of new articles in various news categories

News Category	Count
business	10742
economy	7976
international	3736
politics	20031

3.3.5 Pre-processing of news articles

Webscraped news articles obtained in subsection 3.3.4 are available in html format. They cannot be used in this form in the sentiment extraction process and hence need to be pre-processed. Pre-processing of text involves following steps:

- Converting html text to plain text: Since webscraped text is available in html format, there are html mark up tags present along with the main content. These markup tags are removed so as to get plaintext.

- Removal of stop words: In order to make sentiment analysis process efficient, stop words like “is”, “an”, “shall”, “the”, etc. which are repeated throughout the text and do not convey any sentiment making them irrelevant for sentiment analysis are removed.

- Removal of extra spaces and punctuation marks: Sometimes, text from the websites contain extra spaces in between words or paragraphs, so they are removed. Since, methodology in this thesis follows a simple bag of words approach, punctuation marks are removed.
- Stemming words to their roots: Words like started, starting, etc. are brought to their root word "start" using a process called stemming
- Conversion of text to lower case: In order to reduce sparseness in the text corpus, entire text in the corpus is converted to lower case.

The pre-processed text data is stored as a corpus and used for further analysis.

3.3.6 Extraction and Aggregation of news sentiment

Pre-processing of data not only cleans data by removing unnecessary stop words, punctuation marks, etc. but also converts text data to a plaintext format that can be used in the sentiment extraction process. In this thesis, lexicon-based sentiment analysis method using "Sentometrics" package from R programming language is followed. This package contains Henry's finance dictionary (HE) which is used to identify words with positive and negative sentiment from the the text corpus for each article. Moreover, instead of using unigrams, valence-shifting bigrams approach described in (Ardia et al., 2018) is used. In this approach, a bigram like "not good" gets sentiment score of -1 due to the presence of negator "not" instead of 1 corresponding to "good" in the case of default unigram approach. Further, amplifying valence shifter words such as "very", "highly", etc. strengthen polarized words by 80% whereas deamplifying valence shifter words like "barely", "hardly", etc. downtone them by 80%. Sentiment polarity score is then obtained by using the following formula:

$$S_d = \sum_{i=1}^{Q_d} w_i v_i s_i \tag{3.15}$$

where S_d is the sentiment of a news article d, Q_d is the number of unigrams in document d, v_i is the impact of a valence shifter corresponding to shifting value of the preceding unigram $(i-1)$. No valence shifter or the simple unigrams approach makes $v_i = 1$. If the valence shifter is a negator, then $v_i = -1$. s_i represents the sentiment value attached to unigram i from document d. The weights w_i define the within-document aggregation which for proportional implementation equals $\frac{1}{Q_d}$. Sentiment polarity score[3] thus obtained takes values in the range of -1 to 1. To get aggregate news sentiment, average of news sentiment obtained for articles belonging to all news categories is computed on daily basis. Also, to obtain news sentiment related to four news categories, we separate news sentiment of articles as per news categories viz. business, economy, international, and politics and then find average of news sentiment of articles belonging to their respective news categories on daily basis.

[3] Sentiment polarity score henceforth will be called as news sentiment

After getting news sentiment of each news article, the next step is to aggregate sentiment. In this thesis, aggregation is done in the following ways:

- Aggregation of new sentiment on daily basis: Here aggregation on a given day is done by averaging sentiment polarity of all news articles published on that day. So, the aggregation happens separately for each day in the observation windows.

- Aggregation of news sentiment across news categories on daily basis: Here aggregation of news sentiment for each day in the observation window is done for each news category by taking average of the news sentiment polarity of each news article belonging to that news category. So, in this method the daily aggregation of news sentiment is done for each day in the observation window with respect to the four news categories viz. business, economy, international and politics.

3.3.7 Examination of relationship between news sentiment and returns

Analyses is conducted using stock indices returns as a unit of analyses. Further, daily aggregate news sentiment obtained using the procedure described above is denoted by the variable "Sentiment". Similarly, news sentiment obtained for business, economy, international and politics news categories are denoted by variables "Business", "Economy", "International" and "Politics" respectively. Also, Nifty50 index returns obtained earlier are denoted by the variable "Returns". Returns of the 10 sectoral indices are respectively denoted by the variables Auto, Bank, Energy, FMCG, IT, Media, Pharma, PrivateBank, PSUBank and Realty. Since Nifty50 index is considered as a benchmark index, it represents the Indian stock market and hence in this thesis the "Returns" variable mentioned above as denoting Nifty50 index returns is used as a representative for stock market returns. The relationship between news sentiment variables and the returns of stock market and various sectoral indices are then examined as explained below:

1. Influence of aggregate news sentiment on stock market returns during extreme movement in stock market: Linear regression is used to examine the relationship between the dependent variable Returns and the independent variable Sentiment on day t_0, t_{-2}, t_{-1}, t_{+1}, and t_{+2} days. Strength of the relationship on day t_0 is then compared with that on t_{-2}, t_{-1}, t_{+1}, t_{+2} days.

2. Spillover effect of aggregate news sentiment on stock market returns during extreme movement in stock market: Time-based spillover effect is examined using multiple regression analysis to find out whether news published prior to the day of extreme returns viz. t_{-2} and t_{-1} days have any impact on stock market returns on the day of extreme market movement (t_0).

3. Influence of news sentiment of select news categories on stock market returns during extreme movement in stock market: Multiple linear regression is used to examine the relationship between the dependent variable returns and the independent variables Business, Economy, Politics and International on day t_0, t_{-2}, t_{-1}, t_{+1}, and t_{+2} days. Strength of the relationship on day t_0 is then compared with that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days.

4. Influence of aggregate news sentiment on returns of select sectoral indices during extreme movement in stock market: Each of the 10 sectoral indices returns are examined using linear regression to find out whether they are influenced by aggregate news sentiment on t_{-2}, t_{-1}, t_0, t_{+1}, and t_{+2} days. Regression model on day t_0 is then compared with that on t_{-2}, t_{-1}, t_{+1}, t_{+2} days respectively.

5. Influence of news sentiment of select news categories on returns of select sectoral indices during extreme movement in stock market: Using multiple regression, returns of each of the 10 sectoral indices are examined to find out whether they are influenced by news sentiment from business, economy, politics and international news categories on t_{-2}, t_{-1}, t_0, t_{+1}, and t_{+2} days. Regression model on day t_0 is then compared with regression models on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days respectively.

Chapter 4

Analyses and Findings

This chapter provides details of analyses conducted using methodology described in the previous Chapter 3 and is divided into five sections. Section1 examines influence of aggregate news sentiment on stock market returns using simple linear regression for each day in the observation window. Section 2 examines spillover effect of aggregate news sentiment of previous two days on the returns of day with extreme stock market movement. Section 3 examines influence of news sentiment of business, economy, politics and international news categories on returns of stock market during each day in the observation window. Section 4 examines influence of aggregate news sentiment on returns of various sectoral indices considered in this study during each day of the observation window. Finally, section 6 examines the influence of news sentiment of business, economy, politics and international news categories on returns of various sectoral indices considered in this study during each day in the observation window.

4.1 Relationship between aggregate news sentiment on stock market returns

Regression analysis using Sentiment and Returns variables is conducted to test Hypothesis 1 that there exists a significant relationship between aggregate news sentiment and stock market returns. Results of this analysis shown in Table 4.1 indicate that there is a significant relationship between aggregate news sentiment and market returns on day t_0, $(R^2 = 0.277, F(1,41) = 15.703, p < .01)$. Sentiment was found to be a significant predictor of Returns, $(\beta = 0.526, t(41) = 3.963, p < .001)$. This implies support for Hypothesis 1.

To test Hypothesis 2a, Hypothesis 2b, Hypothesis 2c, and Hypothesis 2d which examine whether relationship on day t_0 is stronger than that on days t_{-2}, t_{-1}, t_{+1}, and t_{+2} respectively, the strength of the relationship in each regression model corresponding to t_{-2}, t_{-1}, t_{+1}, and t_{+2} days is compared with that on day t_0. Table 4.1 shows that while on day t_{-1} the relationship is significant at 10% confidence level $(R^2 = 0.083, F = 3.720, p < .1)$, it is not significant on t_{-2}, t_{+1}, and t_{+2}

Table 4.1: Regression analysis of aggregate news sentiment (Sentiment) and returns of stock market (Returns) within 5-day observation window

	Dependent variable:				
	Returns				
	(t-2)	(t-1)	(t)	(t+1)	(t+2)
	(1)	(2)	(3)	(4)	(5)
Sentiment	0.124	0.288^{*}	0.526^{***}	0.190	0.089
t-statistic	0.801	1.929	3.963	1.241	0.574
p-value	0.428	0.061	0.0003	0.222	0.570
Observations	43	43	43	43	43
R^2	0.015	0.083	0.277	0.036	0.008
Adjusted R^2	-0.009	0.061	0.259	0.013	-0.016
Residual Std. Error (df = 41)	0.013	0.012	0.043	0.018	0.018
F Statistic (df = 1; 41)	0.642	3.720^{*}	15.703^{***}	1.541	0.329
Note:					*p<0.1; **p<0.05; ***p<0.01

days. Also, standardized coefficient of Sentiment, $\beta = 0.288$ on day t_{-1} is smaller as compared to $\beta = 0.526$ on day t_0. This result indicates stronger relationships on day t_0 compared to that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days respectively, thereby supporting Hypothesis 2a, Hypothesis 2b, Hypothesis 2c, and Hypothesis 2d. Using the standardized coefficient from Table 2, the equation for the relationship between Returns and Sentiment on day (t_0) can be written as follows:

$$Returns = 0.526 \times (Sentiment) + 0.043 \tag{4.1}$$

Results of hypothesis testing obtained here are summarized in Table 4.2:

4.2 Spillover effect of aggregate news sentiment on stock market returns

Using multiple regression, spillover effect of aggregate news sentiment of t_{-2} and t_{-1} days is examined on stock market returns of day t_0. Here aggregate news sentiment of days t_{-2}, t_{-1} and t_0 represented as Sent_d1, Sent_d2 and Sent_d3 respectively are taken as independent variables and stock market returns of day t_0 as dependent variable. The result of this analysis is summarized in Table 4.3.

Table 4.3 shows that there is a significant relationship between aggregate news sentiment of

Table 4.2: Summary of results of hypothesis testing related to influence of aggregate news sentiment on returns of stock market

Hypothesis No.	Hypothesis	Remark
Hypothesis 1	There is a significant relationship between aggregate news sentiment and returns of stock market on the day of extreme stock market movement t_0.	Supported
Hypothesis 2a	Relationship between aggregate news sentiment and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-2}.	Supported
Hypothesis 2b	Relationship between aggregate news sentiment and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-1}.	Supported
Hypothesis 2c	Relationship between aggregate news sentiment and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+1}.	Supported
Hypothesis 2d	Relationship between aggregate news sentiment and returns of stock market is stronger on the day of extreme stock market movement as compared to that on day t_{+2}.	Supported

Table 4.3: Regression analysis testing for spillover effect of aggregate news sentiment

	Dependent variable:
	Returns
	t_0
Sent_d1	-0.188
t-statistic	-1.198
p-value	0.239
Sent_d2	-0.344^{*}
t-statistic	-1.950
p-value	0.059
Sent_d3	0.826^{***}
t-statistic	5.291
p-value	0.00001
Observations	43
R^2	0.421
Adjusted R^2	0.377
Residual Std. Error	0.039 (df = 39)
F Statistic	9.462^{***} (df = 3; 39)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

t_{-2}, t_{-1}, and t_0 days and returns of stock market on day t_0, $Adjusted R^2 = 0.377$, $F(3,39) = 9.462$, $p < .01$. Inclusion of the variables Sent_d1 and Sent_d2 alongside Sent_d3 in the regression model improves R^2 to 0.421 compared to 0.277 of the original regression model consisting of only aggregate news sentiment variable on day t_0 (Refer Table 4.1). Variables Sent_d2 $(\beta = -0.344, t(39) = -1.950, p = .059)$ and Sent_d3 $(\beta = 0.826, t(39) = 5.291, p < .001)$ are found to be significant predictors of Returns. Variable Sent_d2 being significant in the model indicates support for Hypothesis 3b. Sent_d1 $(\beta = -0.188, t(39) = -1.198, p = .239)$ is found to be insignificant in the regression model which indicates no support for Hypothesis 3a. Results of hypothesis testing are summarized in the Table 4.4.

Table 4.4: Summary of results of hypothesis testing related to influence of aggregate news sentiment on returns of stock market

Hypothesis No.	Hypothesis	Remark
Hypothesis 3a	There is a spillover effect of aggregate news sentiment of day t_{-2} on stock market returns of day t_0.	Not Supported
Hypothesis 3b	There is a spillover effect of aggregate news sentiment of day t_{-1} on stock market returns of day t_0.	Supported (at 10% confidence level)

4.3 Relationship between news sentiment of news categories on stock market returns

Having found that aggregate news sentiment has a significant relationship with stock market returns on the day of extreme stock market movement and a day prior to it, further examination is done to test whether news sentiment derived from select news categories viz. business, economy, international and politics has any significant relationship with stock market returns on t_{-2}, t_{-1}, t_0, t_{+1}, and t_{+2} days. Results of this regression analysis is reported in Table 4.5. Results on day t_0 indicate that relationship between Returns and independent variables Business, Economy, International and Politics is significant, $Adjusted R^2 = 0.325$, $F(4,38) = 6.065$, $p < .01$. This result indicates support for hypothesis 4 that there is a significant relationship between news sentiment of select news categories and returns of stock market on the day of extreme stock market movement t_0. Table 4.5 further reveals that the variables Business $(\beta = 0.397, t(38) = 3.010, p = .005)$ and Politics $(\beta = 0.543, t(38) = 4.178, p < .001)$ are significant predictors of Returns in the relationship between news sentiment from select news categories and Returns on day t_0. This result indicates support for Hypothesis 4a and Hypothesis

Table 4.5: Regression analysis of Business, Economy, International, and Politics news sentiment and Returns variables in a 5-day observation window

	Dependent variable:				
	Returns				
	t_{-2}	t_{-1}	t_0	t_{+1}	t_{+2}
	(1)	(2)	(3)	(4)	(5)
Business	0.317	0.210	0.397***	0.137	0.033
t-statistic	1.649	1.292	3.010	0.837	0.206
p-value	0.108	0.205	0.005	0.408	0.838
Economy	−0.041	0.118	0.036	0.170	−0.101
t-statistic	−0.262	0.700	0.268	1.078	−0.580
p-value	0.796	0.489	0.790	0.288	0.566
International	0.008	0.018	0.076	−0.067	0.146
t-statistic	0.051	0.107	0.587	−0.421	0.821
p-value	0.960	0.916	0.561	0.677	0.417
Politics	0.263	0.291*	0.543***	0.146	0.084
t-statistic	1.358	1.849	4.178	0.902	0.513
p-value	0.183	0.073	0.0002	0.373	0.612
Observations	43	43	43	43	43
R^2	0.075	0.114	0.390	0.060	0.034
Adjusted R^2	-0.022	0.020	0.325	-0.039	-0.068
Residual Std. Error (df = 38)	0.013	0.012	0.041	0.019	0.019
F Statistic (df = 4; 38)	0.773	1.217	6.065***	0.605	0.329
Note:					*p<0.1; **p<0.05; ***p<0.01

4d. However, the variables International ($\beta = 0.076, t(38) = 0.587, p = 0.561$) and Economy ($\beta = 0.036, t(38) = 0.268, p = 0.790$) are found to be insignificant in this relationship which indicates no support for Hypothesis 4b and Hypothesis 4c. Table 4.5 also shows that relationships between aggregate news sentiment of select news categories and returns on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days are not significant ($p > .05$ *for all relationships*). Also, it can be seen that the standardized coefficients β in the relationship on day t_0 for all variables viz. Business, Economy, International and Politics are more than those on days t_{-2}, t_{-1}, t_{+1}, and t_{+2}. These results indicate that relationship on day t_0 is stronger as compared to that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days respectively and hence provide support for hypothesis 5a, Hypothesis 5b, Hypothesis 5c and Hypothesis 5d. Using standardised coefficients (β) of independent variables Business, Economy, International and Politics on day t_0, equation for Returns can be written as follows:

$$\begin{aligned} Returns &= 0.397 \times (Business) + 0.036 \times (Economy) \\ &+ 0.076 \times (International) + 0.543 \times (Politics) + 0.041 \end{aligned} \tag{4.2}$$

Results of hypothesis testing are summarized as shown in Table 4.6.

Table 4.6: Summary of results of hypothesis testing related to influence of news sentiment from news categories on returns of stock market

Hypothesis No.	Hypothesis	Remark
Hypothesis 4	There is a significant relationship between news sentiment of select news categories and returns of stock market on the day of extreme stock market movement t_0	Supported
Hypothesis 4a	News sentiment from business category has a significant influence on stock market returns on the day of extreme stock market movement t_0	Supported
Hypothesis 4b	News sentiment from economy category has a significant influence on stock market returns on the day of extreme stock market movement t_0	Not Supported
Hypothesis 4c	News sentiment from international category has a significant influence on stock market returns on the day of extreme stock market movement t_0	Not Supported
Hypothesis 4d	News sentiment from politics category has a significant influence on stock market on stock market returns on the day of extreme stock market movement t_0.	Supported
Hypothesis 5a	Relationship between news sentiment of select news categories and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-2}	Supported
Hypothesis 5b	Relationship between news sentiment of select news categories and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-1}	Supported
Hypothesis 5c	Relationship between news sentiment of select news categories and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+1}	Supported
Hypothesis 5d	Relationship between news sentiment of select news categories and returns of stock market is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+2}.	Supported

4.4 Relationship between aggregate news sentiment and returns of sectoral indices

In order to examine influence of aggregate news sentiment on returns of various sectoral indices considered in this study, regression analysis is conducted using aggregate news sentiment (Sentiment variable) and returns of sectoral indices. Results of this analysis is shown in Table 4.7 and Table 4.8. Table 4.7 shows regression model summary of each day in the observation window for all sectoral indices in the form of R^2, f-statistic and p value . Similarly, standardized coefficients (β) of Sentiment for each day in the observation window corresponding to sectoral indices are shown in the Table 4.8. As shown in the Table 4.7, relationship between Sentiment and

Table 4.7: Comparison of regression models involving aggregate news sentiment (Sentiment) and returns of select sectoral indices variables within 5-day observation window

	(t-2)			(t-1)			(t)			(t+1)			(t+2)		
Index	R^2	fstat	p	R^2	fstat	p	R^2	fstat	p	R^2	fstat	p	R^2	fstat	p
Auto	0.056	2.410	0.128	0.062	2.713	0.107	0.186	9.349	0.004	0.068	2.974	0.092	0.000	0.000	0.994
Bank	0.018	0.748	0.392	0.050	2.160	0.149	0.236	12.693	0.001	0.015	0.618	0.436	0.012	0.495	0.486
Energy	0.001	0.025	0.875	0.057	2.486	0.123	0.313	18.654	0.000	0.038	1.616	0.211	0.001	0.055	0.816
FMCG	0.073	3.212	0.080	0.080	3.543	0.067	0.247	13.453	0.001	0.028	1.183	0.283	0.014	0.586	0.448
IT	0.015	0.637	0.430	0.185	9.336	0.004	0.299	17.484	0.000	0.002	0.064	0.802	0.024	1.028	0.317
Media	0.002	0.069	0.794	0.174	8.657	0.005	0.286	16.446	0.000	0.078	3.483	0.069	0.001	0.042	0.838
Pharma	0.013	0.538	0.467	0.120	5.600	0.023	0.254	13.964	0.001	0.014	0.569	0.455	0.029	1.222	0.275
Private Bank	0.010	0.409	0.526	0.014	0.602	0.442	0.264	14.678	0.000	0.013	0.547	0.464	0.010	0.402	0.529
PSU Bank	0.017	0.718	0.402	0.073	3.215	0.080	0.182	9.105	0.004	0.048	2.077	0.157	0.016	0.657	0.422
Realty	0.102	4.678	0.036	0.070	3.101	0.086	0.324	19.628	0.000	0.009	0.369	0.547	0.021	0.897	0.349

Note: *p<0.1; **p<0.05; ***p<0.01

returns of all sectoral indices are found to be significant $(model\, p < 0.05\, for\ all\ sectoral\ indices)$ on day t_0. This result provides support for Hypothesis 6 that there is a significant relationship between aggregate news sentiment and returns of select sectoral indices during extreme movement in stock market. In order to examine strength of the relationships between aggregate news sentiment and returns of sectoral indices day-wise, relationship is checked for its significance in Table 4.7 followed by a comparison of standardized coefficients of the variables in the model shown in the Table 4.8. From Table 4.7, it can be seen that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days, on a confidence level of 5%, except Realty index on day t_{-2} and IT, Media and Pharma indices on day t_{-1}, rest of the relationships are not significant. Further, referring Table 4.8, it can be seen that standardized coefficient (β) of Sentiment corresponding to the regression model related to Realty index on day t_0 is greater (0.569) as compared to that on day t_{-2} which is 0.320. Similarly on day t_0, IT, Media and Pharma indices have standardized coefficient (β) of Sentiment (0.547, 0.535 and 0.504) greater compared to 0.431, 0.418 and 0.347 respectively on day t_{-1}. These results indicate a stronger relationship on day t_0 as compared to that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days respectively. These results thus indicate support for Hypothesis 7a, Hypothesis 7b, Hypothesis 7c and Hypothesis 7d that relationship between aggregate news sentiment and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days. Comparing the standardized coefficients of Sentiment of all sectoral indices in this study (shown in Table 4.8), it can be seen that aggregate news sentiment has the highest influence on returns of Realty index (0.569) while it is the lowest for returns of PSU Bank index (0.426). Results of hypothesis testing obtained above are summarized in Table 4.9.

Table 4.8: Comparison of regression coefficients of aggregate news sentiment (Sentiment) variable in the regression models involving aggregate news sentiment and returns of select sectoral indices within 5-day observation window

		(t-2)	(t-1)	(t)	(t+1)	(t+2)
Index	Statistic	Sentiment	Sentiment	Sentiment	Sentiment	Sentiment
Auto	Coefficient	0.236	0.249	0.431	0.260	-0.001
	t-statistic	1.552	1.647	3.058	1.724	-0.007
	p-value	0.128	0.107	0.004	0.092	0.994
Bank	Coefficient	0.134	0.224	0.486	0.122	0.109
	t-statistic	0.865	1.470	3.563	0.786	0.704
	p-value	0.392	0.149	0.001	0.436	0.486
Energy	Coefficient	0.025	0.239	0.559	0.195	-0.037
	t-statistic	0.159	1.577	4.319	1.271	-0.234
	p-value	0.875	0.123	0.000	0.211	0.816
FMCG	Coefficient	0.270	0.282	0.497	0.167	0.119
	t-statistic	1.792	1.882	3.668	1.088	0.765
	p-value	0.080	0.067	0.001	0.283	0.448
IT	Coefficient	0.124	0.431	0.547	-0.039	0.156
	t-statistic	0.798	3.056	4.181	-0.252	1.014
	p-value	0.430	0.004	0.000	0.802	0.317
Media	Coefficient	0.041	0.418	0.535	0.280	-0.032
	t-statistic	0.263	2.942	4.055	1.866	-0.206
	p-value	0.794	0.005	0.000	0.069	0.838
Pharma	Coefficient	0.114	0.347	0.504	0.117	0.170
	t-statistic	0.734	2.366	3.737	0.754	1.105
	p-value	0.467	0.023	0.001	0.455	0.275
Private Bank	Coefficient	0.099	0.120	0.513	0.115	0.099
	t-statistic	0.640	0.776	3.831	0.740	0.634
	p-value	0.526	0.442	0.000	0.464	0.529
PSU Bank	Coefficient	0.131	0.270	0.426	0.220	0.126
	t-statistic	0.847	1.793	3.017	1.441	0.811
	p-value	0.402	0.080	0.004	0.157	0.422
Realty	Coefficient	0.320	0.265	0.569	0.094	0.146
	t-statistic	2.163	1.761	4.430	0.607	0.947
	p-value	0.036	0.086	0.000	0.547	0.349
Note:						$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4.9: Summary of results of hypothesis testing related to influence of aggregate news sentiment on returns of sectoral indices

Hypothesis No.	Hypothesis	Remark
Hypothesis 6	There is a significant relationship between aggregate news sentiment and returns of sectoral indices on the day of extreme stock market movement t_0	Supported
Hypothesis 7a	Relationship between aggregate news sentiment and returns of sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-2}	Supported
Hypothesis 7b	Relationship between aggregate news sentiment and returns of sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-1}	Supported
Hypothesis 7c	Relationship between aggregate news sentiment and returns of sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+1}	Supported
Hypothesis 7d	Relationship between aggregate news sentiment and returns of sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+2}	Supported

4.5 Relationship between news sentiment of news categories and returns of sectoral indices

Sentiment related to news categories are examined for their influence on returns of select sectoral indices by using multiple regression on t_{-2}, t_{-1}, t_0, t_{+1}, and t_{+2} days. Results of this analysis are given in Table 4.10 and Table 4.11. While Table 4.10 provides day-wise regression model summary for various sectoral indices, Table 4.11 provides summary of standardized coefficients of Business, Economy, International and Politics variables in the regression models involving new sentiment of select news categories and returns of sectoral indices on day t_0.

Referring to Table 4.10, it can be seen that relationships between aggregate news sentiment of select news categories and returns of all sectoral indices on day t_0 are significant ($p < 0.05$ for PSU Bank index and $p < 0.01$ for the rest). This result therefore provides support for Hypothesis 8. Table 4.11 reveals that on day t_0, standardized coefficients of Business variable for regression

Table 4.10: Comparison of regression models involving news sentiment from select news categories (Business, Economy, International and Politics) and returns of sectoral indices within 5-day observation window

	t_{-2}			t_{-1}			t_0			t_{+1}			t_{+2}		
Index	Adj.R^2	F	p	Adj.R^2	F	p	Adj.R^2	F	p	Adj.R^2	F	p	Adj.R^2	F	p
Auto	0.040	1.439	0.240	0.015	1.164	0.342	0.234	4.212	0.006	0.021	1.220	0.318	-0.075	0.271	0.895
Bank	0.053	1.587	0.198	-0.007	0.923	0.461	0.289	5.260	0.002	-0.054	0.466	0.761	-0.067	0.342	0.848
Energy	-0.095	0.088	0.986	0.013	1.137	0.354	0.349	6.641	0.000	0.035	1.384	0.258	-0.069	0.323	0.861
FMCG	0.088	2.017	0.112	0.021	1.228	0.315	0.236	4.247	0.006	-0.040	0.593	0.670	-0.074	0.275	0.892
IT	0.072	1.818	0.146	0.190	3.462	0.017	0.304	5.596	0.001	-0.002	0.978	0.431	-0.053	0.472	0.756
Media	0.083	1.950	0.122	0.156	2.936	0.033	0.293	5.343	0.002	0.231	4.162	0.007	-0.093	0.106	0.980
Pharma	0.094	2.084	0.102	0.123	2.467	0.061	0.277	5.014	0.002	-0.047	0.533	0.712	0.099	2.160	0.092
Private Bank	0.093	2.074	0.103	-0.037	0.623	0.649	0.326	6.083	0.001	-0.076	0.256	0.904	-0.063	0.376	0.824
PSU Bank	-0.019	0.802	0.532	0.033	1.356	0.267	0.177	3.251	0.022	0.070	1.791	0.151	-0.069	0.320	0.863
Realty	0.106	2.244	0.082	0.096	2.116	0.098	0.342	6.447	0.000	-0.081	0.214	0.929	-0.075	0.268	0.897

Note: *p<0.1; **p<0.05; ***p<0.01

models related to Auto, Bank, Energy, FMCG, Private bank, and Realty indices are significant at 1% confidence level and those related to IT, Media, Pharma and PSU Bank indices are significant at 5% confidence level. This result therefore provides support for Hypothesis 8a. On day t_0, standardized coefficient of Business variable corresponding to regression models of all sectoral indices are found to be positively correlated with respective returns of sectoral indices. Also, on day t_0, standardized coefficients of Politics variable for regression models related to all sectoral indices are significant at 1% confidence level. This result therefore provides support for Hypothesis 8d. Standardized coefficient of Politics variable corresponding to regression models of all sectoral indices are found to be positively correlated with respective returns of sectoral indices.

Comparing regression model of all sectoral indices, standardized coefficients of Business

Table 4.11: Comparison of regression coefficients of Business, Economy, International and Politics variables in the regressions involving sentiment from select news categories and returns of sectoral indices on day t_0

Index	Statistic	Business	Economy	International	Politics
Auto	Coefficient	0.400	-0.025	0.015	0.454
	t-statistic	2.846	-0.180	0.108	3.283
	p-value	0.007	0.858	0.914	0.002
Bank	Coefficient	0.401	-0.023	0.096	0.502
	t-statistic	2.960	-0.170	0.728	3.760
	p-value	0.005	0.866	0.471	0.001
Energy	Coefficient	0.414	0.085	0.058	0.549
	t-statistic	3.193	0.655	0.462	4.306
	p-value	0.003	0.516	0.647	0.000
FMCG	Coefficient	0.418	0.100	0.025	0.417
	t-statistic	2.977	0.707	0.183	3.017
	p-value	0.005	0.484	0.856	0.005
IT	Coefficient	0.336	0.107	0.123	0.539
	t-statistic	2.509	0.793	0.943	4.089
	p-value	0.016	0.433	0.351	0.000
Media	Coefficient	0.344	0.117	0.077	0.531
	t-statistic	2.547	0.863	0.586	3.993
	p-value	0.015	0.394	0.562	0.000
Pharma	Coefficient	0.288	0.119	0.058	0.550
	t-statistic	2.104	0.870	0.437	4.091
	p-value	0.042	0.390	0.664	0.000
Private Bank	Coefficient	0.429	-0.016	0.093	0.520
	t-statistic	3.248	-0.118	0.725	4.005
	p-value	0.002	0.907	0.473	0.000
PSU Bank	Coefficient	0.356	0.000	0.085	0.409
	t-statistic	2.438	-0.001	0.599	2.853
	p-value	0.020	0.999	0.553	0.007
Realty	Coefficient	0.376	0.088	0.133	0.551
	t-statistic	2.885	0.672	1.047	4.291
	p-value	0.006	0.506	0.302	0.000

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

variable explains maximum variance (0.429) in case of Private bank index whereas it is minimum (0.288) in case of Pharma index. Similarly, standardized coefficient of Politics variable is found to explain maximum variance (0.551) in case of Realty index and minimum (0.409) in case of PSU Bank index.

Standardized coefficients of International and Economy variables on day t_0 do not show any such significant relationship with returns in regression models of all sectoral indices and hence indicate no support for Hypothesis 8b and Hypothesis 8c.

Referring Table 4.10, it can seen that while on day t_0 all sectoral indices have their relationships significant $(p < 0.05)$, rest of the days t_{-2}, t_{-1}, t_{+1}, and t_{+2} do not have significant relationships for all sectoral indices. Since, only on day t_0, relationships corresponding to all sectoral indices are significant and standardized coefficients of Business and Politics variables in those relationships are also significant(as discussed above), this indicates support for Hypothesis 9a, Hypothesis 9b, Hypothesis 9c and Hypothesis 9d that relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days. Results of hypothesis tests conducted above are summarized in Table 4.12.

Table 4.12: Summary of results of hypothesis testing related to influence of news sentiment from news categories on returns of sectoral indices

Hypothesis No.	Hypothesis	Remark
Hypothesis 8	There is a significant relationship between news sentiment of select news categories and returns of select sectoral indices on the day of extreme stock market movement t_0	Supported
Hypothesis 8a	News sentiment from business category has a significant influence on returns of sectoral indices on the day of extreme stock market movement t_0	Supported
Hypothesis 8b	News sentiment from economy category has a significant influence on returns of sectoral indices on the day of extreme stock market movement t_0	Not Supported
Hypothesis 8c	News sentiment from international category has a significant influence returns of sectoral indices on the day of extreme stock market movement t_0	Not Supported
Hypothesis 8d	News sentiment from politics category has a significant influence on returns of sectoral indices on the day of extreme stock market movement t_0	Supported
Hypothesis 9a	Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-2}	Supported
Hypothesis 9b	Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{-1}	Supported
Hypothesis 9c	Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+1}	Supported
Hypothesis 9d	Relationship between news sentiment of select news categories and returns of select sectoral indices is stronger on the day of extreme stock market movement t_0 as compared to that on day t_{+2}	Supported

Chapter 5

Conclusion

This is the final chapter of the thesis and it wraps up the research by drawing conclusions, highlighting the theoretical contributions and managerial implications. It also reports the limitations of the study and concludes by giving pointers to potential future research areas.

This study emphasized on addressing research questions pertaining to the influence of aggregate news sentiment on returns of stock market as well as returns of sectoral indices. Further, constituents of aggregate news sentiment viz. the news sentiment from business, economy, politics and international categories have also been investigated so as to discover news sentiment of which of these categories influence the returns of stock market and sectoral indices.

In order to ensure that the influence on stock market happens as a result of news sentiment and not because of fundamentals, the study was conducted taking extreme levels of stock market. The idea of taking extreme levels of stock market was considered in accordance with the noise trader theory which suggests that extreme deviations in stock prices from their intrinsic value occur due to sentiment and limits to arbitrage. While noise trader theory focused on investor sentiment formed as a result of the beliefs held by the uninformed investor, this study contributed to this theory by examining aggregate news sentiment as an exogenous factor influencing beliefs held by uninformed investors thereby initiating their participation in the stock market.

The results of this study showed significant influence from aggregate news sentiment on returns of stock market on the day of extreme stock market movement and a day prior to it. Influence on the day prior to extreme stock market movement was found to be weaker compared to the strength on the day of extreme stock market movement. Also there was no significant influence found two days prior and on the next two days after the day of extreme stock market movement. Similar pattern of influence was found on returns of all sectoral indices, however unlike stock market, there was no significant influence found on returns a day prior to the day of extreme market movement. When stock market returns were examined for time-based spillover effect, aggregate news sentiment obtained a day prior to extreme stock market movement showed significant influence on returns of stock market on the day of extreme stock market movement.

On examination of influence of news sentiment from business, economy, international and politics categories, significant influence on returns of stock market and sectoral indices was found only on the day of extreme stock market movement. In these relationships, news sentiment from business and politics were found to be significant predictors of returns of stock market and all sectoral indices whereas news sentiment from economy and international categories were found to be insignificant. On previous two days and next two days after the day of extreme stock market movement, no significant relationship was found between news sentiment from business, economy, international and politics news categories and returns of stock market as well as returns of sectoral indices.

In summary, this study has shown that aggregate news sentiment as a form of noise influences stock market at extreme levels in the short run. These results thus comply with the noise trader theory which also suggests that extreme deviation in stock prices occur due to sentiment and limits to arbitrage. Moreover, this study has shown that apart from news sentiment from business news category, news sentiment from politics news category also influences returns of stock market and sectoral indices.

5.1 Theoretical contribution

The study conducted here attempted to build on to the strand of recent research that suggests news sentiment as one of the important factors that influences stocks' variables like price, returns, volatility, etc. While many of the studies in the past focused on company-specific news sentiment and its impact on individual stocks, this study attempted to study aggregate news sentiment and its influence on portfolio of stocks during extreme movement in stock market. Further, this study also explored what categories of news influences extreme movement in stock market. Some of the theoretical contributions that can be derived from this study are as follows:

5.1.1 Aggregate news sentiment: a manifestation of noise driving stock market to its extreme levels

While influence of company specific news sentiment on individual stocks is well established in the literature (e.g. Cahana, 2013; Ferguson et al., 2011; Heston & Sinha, 2016; Huynh & Smith, 2017; Mo et al., 2016; W. Zhang & Skiena, 2010), the results of this study found that aggregate news sentiment influences returns of stock market and sectoral indices.

When a large number of news articles are published on different events particularly related to the four news categories viz. business, economy, international and politics, market participants have a differing views and beliefs about how these events are going to affect the stock market. Sometimes, an incoming news itself can be a source of a trigger to reinforce their beliefs or change their beliefs. Each news article can be thought of as a carrier of information along with a sentiment component. The sentiment component is induced because of the type of sentiment words used in the news article that can influence the reader's emotion. While information gives the true picture of the event at that point in time, sentiment builds expectations giving rise to a feeling of optimism (pessimism). Since the information and sentiment components are so tied to one another, market participants find it difficult to decipher the degree of information component from sentiment component. Most of the market participants can't filter the information content from the sentiment component. Particularly, the noise traders because of their risk-seeking attitude and eagerness to trade, take positions in the market based on the news they receive without really finding out whether content of news carries enough information component that is relevant for trading at that moment or not. Most of the time, it so happens that they trade on sentiment component rather than the information component that influences their beliefs.

Noise traders' difficulty in trading on information becomes worse when they receive a large number of news articles as they find it even more difficult to pick the appropriate information component to trade on. Instead, they tend to gauge overall sentiment component in the business environment as expressed in the news articles they read. When sentiment polarities of a large number of news articles are the same, this gives rise to a net positive or negative aggregate news sentiment stronger in magnitude to create an impact on the noise traders' beliefs. Sometimes, it may be so strong that it can reinforce beliefs of noise traders, encouraging them to participate in the market because of which stock prices start deviating from their intrinsic value. The deviated prices, along with the aggregate news sentiment can further bring more noise traders in the market or increase the volume of trade from the existing noise traders as it gives confidence to them. This situation is detrimental to inherently risk-averse arbitragers who find it difficult to trade because of their limits to arbitrage. Inability of arbitrageurs to bet against the noisy traders makes the prices deviate to extreme levels. This influence of aggregate news sentiment on stock market during extreme market movement has a close resemblance to the characteristic of noise highlighted by Black (1986) that considers noise induced by a large number of small events often a powerful causal factor than the one caused by a large event. One can also deduce that the aggregate news sentiment influences noise traders' risk by making noise traders to follow their beliefs in the direction of aggregate news sentiment. Aggregate news sentiment can thus be one of the reasons why noise traders' follow their beliefs that do not revert to their mean quickly instead becomes more extreme as suggested by De Long et al. (2003).

While the extant literature shows that researchers like Baker and Wurgler (2007), C. Zhang (2008), etc. using noise trader theory had focused on beliefs of noise traders influencing stock market leading to deviation in stock prices, this thesis has further revealed how these beliefs are influenced by aggregate news sentiment generated from news obtained from business, economy, politics and international news categories creating extreme movement in the stock market.

5.1.2 Short term influence of aggregate news sentiment on stock market returns

According to the results of this study, aggregate news sentiment influences stock market returns on the day of extreme market movement. But, in order to verify whether similar results are obtained for other days as well, influence created by aggregate news sentiment on t_{-2}, t_{-1}, t_{+1}, and t_{+2} days is compared with that on day t_0; day t_0 being the day of extreme returns. Results indicate that except day t_{-1}, there is no significant influence created by aggregate news sentiment on days t_{-2}, t_{+1}, and t_{+2} days. Also, strength of the influence created by aggregate news sentiment on day t_{-1} is relatively lower compared to that on the day of extreme returns t_0. This shows that aggregate news sentiment has relatively weaker influence on stock market on day t_{-1} as compared to that on day t_0. Thus, it means that aggregate news sentiment has a very short term impact that begins on the day prior to the day of extreme returns and ends on the day of extreme returns.

The results reported in this thesis are similar to the ones reported by Ferguson et al. (2011), Gidofalvi (2001), Heston and Sinha (2016), W. Zhang and Skiena (2010) in terms of the short term influence of news sentiment on returns of stocks. While in this thesis, aggregate news sentiment was found to have influence on stock market returns on the day of extreme returns and relatively weaker influence a day prior to it, Ferguson et al. (2011) reported significant influence of both positive and negative news sentiment on returns for companies in UK on the day of news release and weaker influence only from negative news sentiment on the next day. Heston and Sinha (2016) also found predictability of stock returns using daily news sentiment for a short period of one to two days.

The results of this thesis also conform with the results reported by W. Zhang and Skiena (2010) where they found a significant correlation between polarity of news and returns on the day of news release. Further, they also found that news get reflected in stock prices within a day of their release thereby indicating a short term influence as compared to relatively long term influence for blogs to get reflected in prices.

The short term influence of news sentiment is also reported by Gidofalvi (2001) who found that news published on stocks have an impact 20 minutes before and 20 minutes after their news is published but with low predictive power. One of the reasons he cites for low predictive power is that while important news is reported repeatedly in different news articles by different agencies, presumably only the first article has an influence on the stock prices. In this thesis, the observation window is of five days compared to 40 minutes interval in his study. So, there is a higher possibility of some news articles being repeated or mere updates of the already published news article and hence the novelty factor is only applicable for those news articles which have published content about events first time and are not updates. This could also be the reason why the various regression models related to stock market and sectoral indices in this thesis have a lower R^2s or low predictive power.

This thesis conforms with the evidences provided by Ferguson et al. (2011), Gidofalvi (2001), Heston and Sinha (2016), W. Zhang and Skiena (2010) for short term influence of news sentiment. However, evidences on influence of news sentiment over longer periods has also been revealed by Kräussl and Mirgorodskaya (2013), Uhl et al. (2015), etc. as discussed in the literature review.

The explanation for extreme movement of stock market under the influence of aggregate news sentiment has already been given earlier in the subsection 5.1.1. Having achieved the extreme levels, the stock market doesn't sustain it for long; within a day it starts correcting and then showing no significant influence of aggregate news sentiment on the day following this extreme levels i.e. on day t_{+1}. In other words, it means that content of the news articles on day t_{+1} is either devoid of feeling of optimism (pessimism) or it is opposite in sentiment polarity to the one on day t_0. The aggregate news sentiment devoid of optimism (pessimism) can happen due to mismatch of sentiment polarities of various news articles received during that period (day) which cancel out one another without giving much confidence to noise traders to trade in the same direction. Also, even if the news articles on the day t_{+1} are written on the same events which led to extreme market movement on day t_0, the aggregate news sentiment may not be that influential having lost novelty or surprise element in the news because of the previous day's coverage. Aggregate news sentiment going in the opposite direction on the day t_{+1} may start forcing noise traders to reverse direction of their trade resulting in booking profits (losses) but won't be enough to give them confidence to do it heavily unless they change their beliefs which in a short period of one day may be a rare case.

5.1.3 Categories of news that influence stock market returns

Out of the four news categories examined, news sentiment from business and politics categories influence stock market in the short term. Since news from business category comprises of news related to stocks, industry and market, this finding is similar to other studies like Cahana (2013), Ferguson et al. (2011), Ranco et al. (2015), W. Zhang and Skiena (2010), etc. wherein they reported influence of company specific news sentiment on stock's variables like price, returns, etc. Moreover, news sentiment from politics category showing a significant contribution proves that investors get influenced by sentiments emanating from political situations, policies, etc. However this is not the case with the news sentiment from economy and international news categories suggesting that they do not create enough news sentiment for noise traders to participate and drive stock market to extreme levels.

Even though one of the objectives of this thesis was to find out which categories of news influence stock market and the results are obtained here with respect to that, further question that needs to be answered is "what characteristics of news in a certain news category makes difference in their influence on returns of stock market or sectoral indices?" This may require a separate research altogether, but considering the categories of news, literature review in this thesis and knowledge gained through this research, it is possible to provide here some logical arguments which may be useful for further research. One of the ways to understand the difference in the influence from news sentiment of different news categories is by referring to the study by Leinweber and Sisk (2011), wherein they identify four important characteristics of news important for sentiment analysis viz. news intensity, relevance, probability of a news item being positive, negative, or neutral in tone and novelty. Although a quantitative study is not done in this thesis with respect to all these characteristics, some of the logical arguments with respect to each news category can be given as follows:

- Business news category: Business news category consists of news related to stocks, market and industry. Market participants are more inclined to reading news from this category as it covers news on events related to company, industry and stock markets. So, relevance of news is quite high in this category of news. Data collected in this thesis suggests that events in this category of news gets wide coverage in terms of number of news articles written on them (Table 3.5 showing 10,742 news item which is the second highest in terms of percentage (25.28%) of the total news filtered in this thesis. So, the news intensity for business category of news is quite high. The very fact that this study has established relationship between news sentiment obtained from this news category and returns of stock market and sectoral indices, it given evidence of news possessing sentiment(tone). Although, this thesis has not specifically focused on novelty, some of the news articles gathered in this thesis is likely to possess novelty characteristic. With most of the characteristics of news discussed here with respect to business news category are favorable, news sentiment of business news category is found to be significant in its influence on returns of stock market and sectoral indices.

- Politics news category: Politics news category consists of news related to political events, policy decisions by government, government formation, criticisms from opposition parties, etc. This type of news is often sensationalized and is covered by news media giving regular updates. A large number of events in this news category are also highly unpredictable which may result in such news having high novelty. The volume of news published is large as can be seen from the data collected in this thesis which suggests that events in this category of news gets wide coverage in terms of number of news articles written on them; Table 3.5 showing 20,031 news item which is the highest in terms of percentage (47.15%) of the total news filtered in this thesis. Political news gets attention from all sections of the society as some of the events have major influence on country's macro environment including stock market. Hence, market participants keenly observe political events and build expectations either in favor of stock market or against it. This makes politics news category highly relevant for stock market. Study conducted in this thesis has already shown the presence of sentiment in the news related to political news category. Since most of the characteristics of the news related to politics news category are highly favorable, politics news category is found to be significant in its influence on the returns of stock market and sectoral indices.

- Economy news category: Economy news category consists of news related to economic policies, announcements about economic indicators like Gross Domestic Product (GDP), Consumer Price Index (CPI), Interest rates, unemployment, etc., economic events and other economic decisions taken by RBI (reserve Bank of India) and Government. Except news related certain economic events and policies, most of the other economic news are often on the expected lines and often discussed by experts on various electronic media and hence known to general public in advance. Moreover, in this news category, a large number of news articles are released along with the quantitative data in the form of indicators which is given more prominence by experts and that makes the sentiment component of the news articles weaker especially in the short run. Considering these characteristics of news related to economy news category, the novelty component in the news gets diminished. The volume of news in this news category is usually less as can be seen from the data collected in this thesis which suggests that events in this category are either few or do not get wide coverage as compared to events related to other news categories in terms of number of news articles written on them; Table 3.5 showing 7,976 news item which is the relatively less in terms of percentage (18.77%) of the total news filtered in this thesis. Even though news related to economy possess content having sentiment, the other characteristics do not support it to have a significant influence on the returns of the stock market or sectoral indices in the short run. One may have to examine the influence of news related to economy in medium and long term may be due to the long terms expectations of the market participants.

- International news category: International news category consists of news from general domain and its coverage ranges from geopolitical situations, international trade agreements, visits of eminent personalities across boundaries, events like terrorist attacks, wars, natural calamities, major cultural and sports events across the globe,etc. While some type of international news like geo-political situations, wars, international trade agreements, etc. do make an impact on global landscape, the impact happens over a period of time unless a country is directly involved in it. Other general type of news from global events may not have much relevance domestic stock market. Time zone of different countries where an event happened and the timing of stock market trading where the impact is observed in the study may be different which make the news lose novelty. The volume of news in this news category is usually very less as can be seen from the data collected in this thesis which suggests that events in this category of news gets low coverage in terms of number of news articles written on them as usually media in a particular country is more likely to give more attention to domestic events compared to global ones unless they are highly significant in the global landscape; Table 3.5 showing 3,736 news item resulting in the lowest percentage (8.79%) of the total news filtered in this thesis. Probably because of these reasons, international news category has not shown a significant influence on the returns of stock market or sectoral indices in the short run.

5.1.4 Spillover effect of aggregate news sentiment

In this study spillover effect of aggregate news sentiment of day t_{-1} was found on stock market returns of day t_0 which is the day of extreme market movement. Smales (2016) also reported a similar result of their study wherein they found that the current period and lagged news sentiment (t_{-1}, t_{-5}, t_{-20}, t_{-60}, and t_{-250}) have a significant positive (negative) relationship with stock returns. However, as compared to these results, Mo et al. (2016) reported lag-5 (t_{-5}) effect of news sentiment on market returns and conversely lag-1(t_{-1}) effect of market returns on news sentiment. Audrino and Tetereva (2017) also reported presence of spillover effect during periods of crisis like financial instability which is a period of extreme market movement. However, their study revealed spillover effect of news sentiment across returns of different sectors.

The result of this study thus provides an evidence for how aggregate news sentiment is reflected in prices. The pattern here indicates that it takes a short while (approximately one day) for sentiment to gain momentum as can be seen from weak impact of previous day's aggregate news sentiment on returns of stock market on the day of extreme movement in the market. This influence of aggregate news sentiment is having a pattern similar to the one explained by Hong and Stein (1997) for underreaction and overreaction. While their model gives importance for market participants called "news watchers" who tend to initiate the direction of movement which is further exemplified by the "momentum traders" by watching the price move in a particular direction, the study conducted in this thesis considered both "news watchers" and "momentum traders" as "noise traders".

5.1.5 Sectoral indices follow reaction of stock market to aggregate news sentiment

While various studies on news sentiment as discussed in the chapter on literature review mostly focused on stock-specific news sentiment, this thesis has provided an evidence for the influence of aggregate news sentiment on the returns of various sectoral indices. This is a significant contribution towards understanding how aggregate news sentiment influences returns of various sectoral indices in the short run considering that each sectoral index is constructed to examine the performance of that sector. While it is known that the economic activity in a particular sector is bound to influence that sector index, there is also a spillover effect of the aggregate news sentiment that is reflected on the returns of all the sectoral indices when the stock market itself is significantly influenced by aggregate news sentiment. This can be understood from the results of the study wherein it is found that along with the returns of stock market, the returns of sectoral indices are also influenced by aggregate news sentiment on the day of extreme movement in stock market. However, unlike stock market, in the case of all sectoral indices there was no significant influence found on the returns a day prior to the day of extreme stock market movement. This is an indication of the aggregate news sentiment beginning to influence returns of stocks particularly the ones representing the stock market (Nifty50 index) a day prior to the extreme market movement and then this influence of aggregate news sentiment spills over to other stocks from different sectors. While findings of Baker and Wurgler (2007) reveal that due to difficulty and subjectivity in determining true value of speculative stocks, they tend to be more sensitive to investors' sentiment compared to the stocks with large capitalization having long earnings history, tangible assets, and stable dividends, the results of this study provides evidence for influence of aggregate news sentiment on large cap stocks which are part of stock market index viz. the Nifty50 index. One of the reasons why large cap stocks are also affected by sentiment is because of the noisy signals in the newsflow which are expressed as aggregate news sentiment in the business environment whose degree of influence cannot be valued easily by the market participants. So, rather than depending on the individual characteristics of the stocks, the noise traders rely on the noisy signals in the form of aggregate news sentiment and tries to gauge their impact on stocks using their beliefs. However, they still need to decide which stock(s) to buy or sell. This problem of selection is more prominent when buying stocks than selling as there are a large number of stocks to choose from during buying compared to a limited number of stocks to sell in the portfolio. This problem has been studied by Barber and Odean (2008) and found that individual investors who are generally noise traders tend to buy stocks which catch their attention and this attention-driven buying happens irrespective of whether the stock is a large cap stock or a small cap stock. Although, their study doesn't find this attention-driven effect applicable generally when investors sell their stocks due to limited number of stocks in their portfolio, noise traders will experience disposition effect proposed by Shefrin and Statman (1985) which will encourage them to sell their winner stocks early and ride on losers stocks. Particularly the case of selling stocks early would be more prominent in case there is a negative news sentiment about the stock or the negative aggregate news sentiment affecting the entire stock market. With attention-driven effect and disposition effect finding prominence in the selection of stocks for buying and selling under the influence of aggregate news sentiment, it is important to find out what type of stocks grab attention. News articles often report attention-grabbing events and their influence on global as well as domestic stock market. Since large cap stocks constituting the stock market (Nifty50 index) are highly tracked stocks by institutional as well as retail investors, they will feature more in the analysts' columns, news media as well as social media compared to midcap and small cap stocks. Since they are in the limelight, they will grab attention of noise traders before other stocks. Hence, aggregate news sentiment will influence stock market returns to begin with and this influence will spill over to other stocks including the ones from various sectoral indices when there is extreme movement in the stock market.

Another contribution resulting from examination of influence of aggregate news sentiment on the returns of sectoral indices is that compared to other sectors, aggregate news sentiment significantly influences returns of Realty sector the most whereas this influence is the least for returns of PSU Bank sector. In comparison to this evidence, Borovkova and Lammers (2017) found that news sentiment influences the returns of financial sector most significantly compared to other sectors.

In summary, this research contributes to the existing literature in the area of behavioral finance by studying aggregate news sentiment as a general case. This study finds influence of aggregate news sentiment stronger when stock market returns reach extreme levels due to the participation of noise traders.

5.2 Managerial Implications

This study contributes to the existing literature on noise trader theory that considers sentiment an important factor for explaining deviation in stock prices from their intrinsic value. With investors having access to information, judicious use of news sentiment in their strategies can improve their performance in the market. This research work has following managerial implications:

1. A large number of market participants track the movement of stock market indices like BSE's Sensex and NSE's Nifty 50 index. This study has revealed that the influence of aggregate news sentiment on returns of stock market (Nifty50 index) is short-lived; not even lasting for a day after extreme returns. In this situation, investors should be cautious when stock market is volatile and achieves extreme levels.
2. Findings of this research provide evidence for influence of aggregate news sentiment on stock market during extreme returns. This implies that there are noise traders active in the market who trade on noisy signals obtained from news media and drive stock market to extreme level. These signals are found to be emanating from not only business but also politics news category. So, arbitrageurs should be cautious when there is an extreme pessimism or optimism in the market due to events related to business and politics.
3. Traders in the derivative segment can design strategies in the Nifty50 index futures and options by studying business and political events lined up in the near future and trade with caution.
4. Since influence of news sentiment is high on the day of extreme returns, traders who are interested in taking intraday positions in the market can take advantage of the events related to business and politics if they can gauge sentiment of the events and trade cautiously.
5. Apart from India's leading benchmark indices like Nifty50 and Sensex, retail investors as well as fund managers keep track of movement in sector specific indices. They are always on a look out for periods where certain sectors perform better than the others so that they can use sector rotation strategies. In this thesis it is found that when stock market deviates to extremes, the aggregate news sentiment also has a significant influence on returns of all

sectoral indices considered in this study. However, there is not much difference seen in this influence across various sectors. So, it is not possible for retail investors and fund managers to use the short term influence of aggregate news sentiment for sector rotation strategies in the funds that they manage.

6. Since news sentiment influences stock market, it shows how sensitive investors are to the news published in India. So, news reporting should be done responsibly by using appropriate sentiment words especially when covering events having negative impact on the public.

5.3 Limitations of the Study

The study conducted in this thesis suffers from following limitations:

1. Henry's Finance dictionary has only 105 positive words and 85 negative words which poses a limitation on this study if news articles contain sentiment words which do not exist in this dictionary.

2. This study is limited by the limitations associated with document level sentiment analysis. For instance, document level sentiment analysis do not capture sarcasm in sentences of news articles and hence may not reveal the correct sentiment associated with it. Also, it does not take into account the context in which a sentiment word is being used. For instance, a "low" with regards to "cost" has positive sentiment but "low" in the context of "sales" has negative sentiment.

3. The study has a mix of news articles from business, politics, economics and international categories, some of which are published for the first time covering new events but some are also updates of the old news articles which have already covered a particular event earlier. According to Leinweber and Sisk (2011), novelty in news is one of the important characteristics. This thesis does not differentiate this aspect of online content and therefore gives same importance to both news articles that provide information on a new event and the ones that are just updates.

5.4 Future research

Even though this research follows a parsimonious approach to study aggregate news sentiment and its influence on the stock market, it has given some encouraging results for further research. Some pointers for further research are given below:

- Influence of aggregate news sentiment on different types of portfolio of stocks may be performed during extreme returns in market. For instance, one may select portfolios based on market capitalization, beta, etc.

- One may drill down news categories further to get named entities and their influence on stock market. For instance, one may examine news sentiment related to the ruling party or the opposition party and examine their influence on the stock market.
- Comparative study may be performed on the influence of positive and negative aggregate news sentiment on the stock market returns during extreme returns in stock market.
- Rather than obtaining aggregate news sentiment from extreme returns of Nifty50 index, one may obtain aggregate and category-wise news sentiment from extreme returns of respective sectoral indices and similar study may be conducted to find out influence of aggregate and category-wise news sentiment on their returns.
- This thesis has revealed that news sentiment from business and politics news categories are having significant influence on the returns of stock market as well as sectoral indices as compared to the one from economy and international news categories. However, a thorough investigation is required to reveal what characteristics of the news published related to the four news categories that make the news sentiment from these news categories different in terms of their influence on the returns of stock market and sectoral indices.
- Instead of taking returns as a response variable, study may be conducted to examine influence of aggregate news sentiment and news sentiment from various news categories on other stock market variables like trading volume, liquidity, etc.

This study has opened vistas for further research in the area of aggregate news sentiment and its influence on stock markets in the short term. It has also brought extreme deviations in stock market into limelight through the lens of aggregate news sentiment and in accordance with the noise trader theory.

Bibliography

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, *59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2017). A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level, (201), 1–12.

Aravind, M. (2016). Contrarian and Momentum Strategies: An Investigation with reference to Sectoral Portfolios in NSE. *Nmims Management Review*, *29*(November), 102–117.

Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2018). The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment. *SSRN Electronic Journal*, (2018). https://doi.org/10.2139/ssrn.3067734

Audrino, F., & Tetereva, A. (2017). Sentiment Spillover Effects for US and European Companies. *SSRN Electronic Journal*, 1–42. https://doi.org/10.2139/ssrn.2957581

Augustyniak, L., Szymánski, P., Kajdanowicz, T., & Tuliglowicz, W. (2016). Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy*, *18*(1), 1–29. https://doi.org/10.3390/e18010004

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2)arXiv arXiv:1011.1669v3, 129–151. https://doi.org/10.1257/jep.21.2.129

Bandopadhyaya, A., & Jones, A. L. (2006). Measuring investor sentiment in equity markets. *Journal of Asset Management*, *7*(3-4)arXiv arXiv:1011.1669v3, 208–215. https://doi.org/10.1057/palgrave.jam.2240214

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, *21*(2), 785–818. https://doi.org/10.1093/rfs/hhm079

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*(3), 307–343. https://doi.org/10.1016/S0304-405X(98)00027-0

Bikhchandani, S., & Sharma, S. (2011). *Herd Behavior in Financial Markets* (No. *8*). https://doi.org/10.1088/1751-8113/44/8/085201

Birz, G., & Dutta, S. (2016). U.S. Macroeconomic News and International Stock Prices: Evidence from Newspaper Coverage. *Accounting and Finance Research*, *5*(1), 247–252. https://doi.org/10.5430/afr.v5n1p247

Black, F. (1986). Noise. *The Journal of Finance*, *41*(3), 528–543.

Boiy, E., Hens, P., Deschacht, K., & Moens, M.-f. (2007). Automatic Sentiment Analysis in On-line Text Concepts of Emotions in Written Text Concept of Emotions. *Electronic Publishing*, (June), 349–350. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Automatic+sentiment+analysis+in+on-line+text#0

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Borovkova, S., & Lammers, P. (2017). Sector News Sentiment Indices. *SSRN Electronic Journal*, (November). https://doi.org/10.2139/ssrn.3080318

Brown, K., & Harlow, W. (1988). Market overreaction: Magnitude and intensity. *The Journal of Portfolio Management*, *14*(2), 6–13. https://doi.org/10.3905/jpm.1988.409137

BSE. (2016). Company overview. https://doi.org/10.3109/9781439810828-14

Cahana, S. F. (2013). Media Sentiment, Investor Sentiment, and Stock Price Sensitivity to Earnings. *Management and Service Science (MASS), 2011 International Conference*, 1–4.

Chari, S. G., Hegde Desai, P., & Borde, N. (2017). A review of literature on short term overreaction generated by news sentiment in stock market. *Anushandhan*, *7*(1), 12–21.

Chowdhury, S. G., Routh, S., & Chakrabarti, S. (2014). News Analytics and Sentiment Analysis to Predict. *International Journal of Computer Science and Information Technologies*, *5*(3), 3595–3604.

Cutler, D. M., Poterba, J. M., Summers, L. H., M., P. J., & H., S. L. (1989). What moves stock prices? *The Journal of Portfolio Management*, *15*(3), 4–12. https://doi.org/10.3905/jpm.1989.409212

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and investor security market under-and overreactions. *Journal of Finance*, *53*(6), 1839–1886.

Davis, M. H. A. (2006). Louis Bachelier's " Theory of Speculation ", (2001), 1–23. https://f-origin.hypotheses.org/wp-content/blogs.dir/1596/files/2014/12/Mark-Davis-Talk.pdf

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., & Shleifer, A. (2003). Noise Trader Risk in Financial Markets. *Inefficient Markets*, *98*(4), 28–52. https://doi.org/10.1093/0198292279.003.0002

Dhankar, R. S., & Maheshwari, S. (2014). A Study of Contrarian and Momentum Profits in Indian Stock Market. *International Journal of Financial Management*, *4*(2), 40–54. http://www.publishingindia.com

Drury, B., Torgo, L., & Almeida, J. J. (2012). Classifying news stories with a constrained learning strategy to estimate the direction of a market index. *International Journal of Computer Science and Applications*, *9*(1), 1–22.

Edwards, R. D., Magee, J., & Bassetti, W. H. C. (2018). *Technical Analysis of Stock trends* (11th ed.). CRC Press.

Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives*, *15*(4)arXiv arXiv:1011.1669v3, 157–168. https://doi.org/10.1257/jep.15.4.157

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, *38*(1), 34–105.

Fama, E. F. (1970). Efficient Capital Markets-A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2)arXiv arXiv:1011.1669v3, 383–417. https://doi.org/10.2307/2329297

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance*, *31*(1)arXiv arXiv:1011.1669v3, 139–141. https://doi.org/10.1111/j.1540-6261.1976.tb03204.x

Fenzl, T., & Pelzmann, L. (2012). Psychological and Social Forces Behind Aggregate Financial Market Behavior. *The Journal of Behavioral Finance*, *13*, 56–65. https://doi.org/10.1080/15427560.2012.655383

Ferguson, N. J., Michael, J., Lam, H. Y. T., & Philip, D. (2011). Media Sentiment and UK Stock Returns. *Durham University*, *86*(0).

Geetha, C., Chong, V., Mohidin, R., & Chandran, V. V. (2011). The Relationship Between Inflation And Stock Market : Evidence From Malaysia , United States And China. *International Journal of Economics and Management Sciences*, *1*(2), 1–16.

Giachanou, A., & Crestani, F. (2016). Like It or Not : A Survey of Twitter Sentiment Analysis Methods, *49*(2).

Gidofalvi, G. (2001). Using New s Articles to Predict Stock Price Movements, In *Cit'19 international conference on integrated transport 2019 - impacts of disruptive technologies on travel and transport view*.

Gilson, R. J., & Kraakman, R. H. (2003). The Mechanisms of Market Efficiency Twenty Years Later. *The Journal of Corporation Law*, 715–744.

Guha, B., Dutta, A., & Bandyopadhyay, G. (2016). Measurement of Risk Vs Return of Indian Sectoral Indices. *Journal of Advanced Management Science*, *4*(July 2018), 106–111. https://doi.org/10.12720/joams.4.2.106-111

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, *2*(3), 153–170. https://doi.org/10.1016/j.jfds.2017.02.001

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, *45*(4), 363–407. https://doi.org/10.1177/0021943608319388

Henry, E., & Leone, A. J. (2011). Measuring Qualitative Information in Capital Markets Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1470807

Heston, S. L., & Sinha, N. R. (2016). News versus Sentiment: Predicting Stock Returns from News Stories. *Finance and Economics Discussion Series*, *2016*(048), 1–35. https://doi.org/10.17016/feds.2016.048

Hong, H., & Stein, J. (1997). A unified theory of underreaction, momentum trading and overreaction in asset markets. *The Journal of Finance*, *54*(6). http://www.nber.org/papers/w6324

Humpe, A., & Macmillan, P. (2009). Can macroeconomic variables explain long-term stock market movements? A comparison of the US and Japan. *Applied Financial Economics*, *19*(2), 111–119. https://doi.org/10.1080/09603100701748956

Huynh, T. D., & Smith, D. R. (2017). Stock Price Reaction to News: The Joint Effect of Tone and Attention on Momentum. *Journal of Behavioral Finance*, *18*(3), 304–328. https://doi.org/10.1080/15427560.2017.1339190

Isidore, R., Christie, P., & Christoe, P. (2018). Fundamental Analysis versus Technical Analysis: A comparative review. *International Journal of Recent Scientific Research*, *9*(1 (B)), 23009–23013. https://doi.org/10.24327/ijrsr

Jareño, F., & Negrut, L. (2016). US stock market and macroeconomic factors. *Journal of Applied Business Research*, *32*(1), 325–340. https://doi.org/10.19030/jabr.v32i1.9541

Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, *6*(2-3)arXiv arXiv:1011.1669v3, 95–101. https://doi.org/10.1016/0304-405X(78)90025-9

Jivani, A. (2007). A Comparative Study of Stemming Algorithms. *October*, *2*(2004), 1930–1938.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185. https://doi.org/10.1016/j.irfa.2014.02.006

Kirkpatrick, C., & Dahlquist, J. (2011). *Technical Analysis: The Complete Resource for Financial Market Technicians.* (2nd). Pearson Education, Inc. https://doi.org/10.1002/9781118531785.ch1

Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination arXiv 1507.00955. http://arxiv.org/abs/1507.00955

Kräussl, R., & Mirgorodskaya, E. (2013). *News Media, Framing and Investor Behavior* (June), Center for Financial Studies (CFS). https://doi.org/10.2139/ssrn.2246295

Leinweber, D., & Sisk, J. (2011). Event-Driven Trading and the New News. *The Journal of Portfolio Management*, *38*(1), 110–124. https://doi.org/10.3905/jpm.2011.38.1.110

Liew, V. K. S., & Rowland, R. (2016). The effect of Malaysia general election on stock market returns. *SpringerPlus*, *5*(1). https://doi.org/10.1186/s40064-016-3648-5

Loughran, T., & Mcdonald, B. (2011). When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks Journal of Finance , forthcoming. *Journal of Finance*, *66*(1)arXiv arXiv:1011.1669v3, 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, *35*(March)arXiv 0022-0515, 13–39. https://doi.org/10.2307/2729691

Malandrakis, N., Kazemzadeh, A., Potamianos, A., & Narayanan, S. (2013). SAIL: A hybrid approach to sentiment analysis. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, *2*(SemEval), 438–442.

Malkiel, B. G. (1973). *A random walk down wall street*. http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *arXiv preprint*arXiv 1112.1051, 10. http://arxiv.org/abs/1112.1051

Mcwilliams, A., & Siegel, D. (1997). Event studies in management research: Theoretical and empirical issues. *Academy of Management Journal*, *40*(3), 626–657. https://doi.org/10.2307/257056

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Meier, J. H., Esmatyar, W., & Frost, R. (2018). The predictive power of the sentiment of financial reports. *CEUR Workshop Proceedings*, *2104*, 30–44.

Mian, M., & Sankaraguruswamy, S. (2012). Investor sentiment and stock market response to earnings news. *Accounting Review*, *87*(4), 1357–1384. https://doi.org/10.2308/accr-50158

Mo, S. Y. K., Liu, A., & Yang, S. Y. (2016). News sentiment to market impact and its feedback effect. *Environment Systems and Decisions*, *36*(2), 158–166. https://doi.org/10.1007/s10669-016-9590-9

Nagar, A., & Hahsler, M. (2012). Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams. *International Proceedings of Computer . . .*, *47*(Iccts), 91–95. https://doi.org/10.7763/IPCSIT.2012.V47.18

Nigam, K., Lafferty, J., & Mccallum, A. (1999). Using Maximum Entropy for Text Classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 61–67. https://doi.org/10.1.1.63.2111

NSE. (2015). NSE - National Stock Exchange of India Ltd. https://www1.nseindia.com/global/content/about_us/about_us.htm

Oler, D. K., Harrison, J. S., & Allen, M. R. (2008). The danger of misinterpreting short-window event study findings in strategic management research: An empirical illustration using horizontal acquisitions. *Strategic Organization*, *6*(2), 151–184. https://doi.org/10.1177/1476127008090008

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*arXiv arXiv:0906.0612, 1320–1326. https://doi.org/10.1371/journal.pone.0026624

Pang, B., Lee, L. L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 6-7, 2002, Philadephia, Pennsylvania, USA*arXiv 0205070v1, 79–86. https://doi.org/10.3115/1118693.1118704

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trendsőin Information Retrieval*, *2*arXiv 0112017, 91–231. https://doi.org/10.1561/1500000001

Petrusheva, N., & Jordanoski, I. (2016). Comparative analysis between the fundamental and technical analysis of stocks. *Journal of Process Management. New Technologies*, *4*(2), 26–31. https://doi.org/10.5937/JPMNT1602026P

Piccoli, P., Chaudhury, M., Souza, A., & da Silva, W. V. (2017). Stock overreaction to extreme market events. *North American Journal of Economics and Finance*, *41*, 97–111. https://doi.org/10.1016/j.najef.2017.04.002

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications*, 1–10. https://doi.org/10.1007/1-4020-4102-0_1

Prechter, R. R., & Parker, W. D. (2007). The Financial/Economic Dichotomy in Social Behavioral Dynamics: The Socionomic Perspective. *Journal of Behavioral Finance*, *8*(2), 84–108. https://doi.org/10.1080/15427560701381028

Quaye, I., Mu, Y., Abudu, B., & Agyare, R. (2016). Review of Stock Markets' Reaction to New Events: Evidence from Brexit. *Journal of Financial Risk Management*, *05*(04), 281–314. https://doi.org/10.4236/jfrm.2016.54025

Ranco, G., Aleksovski, D., Caldarelli, G., Grar, M., & Mozeti, I. (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE*, *10*(9)arXiv 1506.02431, 1–21. https://doi.org/10.1371/journal.pone.0138441

Schumaker, R. P., & Chen, H. (2006). Textual analysis of stock market prediction using financial news articles. *Association for Information Systems - 12th Americas Conference On Information Systems, AMCIS 2006*, *3*(2006), 1422–1430.

Shefrin, H., & Statman, M. (1985). The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. *The Journal of Finance*, *40*(3), 777–790. https://doi.org/10.1111/j.1540-6261.1985.tb05002.x

Shi, Y., Ho, K. Y., & Liu, W. M. (2016). Public information arrival and stock return volatility: Evidence from news sentiment and Markov Regime-Switching Approach. *International Review of Economics and Finance*, *42*, 291–312. https://doi.org/10.1016/j.iref.2015.12.003

Shleifer, A., & Summers, L. H. (1990). The Noise Trader Approach to Finance. *Journal of Economic Perspectives*, *4*(2), 19–33. https://doi.org/10.1257/jep.4.2.19

Smales, L. A. (2016). Time-varying relationship of news sentiment, implied volatility and stock returns. *SSRN Electronic Journal*.

Stone, P. J., & Hunt, E. B. (1963). A Computer Approach to Content Analysis: Studies Using the General Inquirer System, In *Proceedings of the may 21-23, 1963, spring joint computer conference*. https://doi.org/10.1145/1461551.1461583

Tetlock. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Thaler, R., & De Bondt, W. F. M. (1985). Does the Stock Market Overreact? *The Journal of Finance*, *40*(3), 793–805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

Timmermann, A., & Granger, C. W. J. J. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, *20*(1), 15–27. https://doi.org/10.1016/S0169-2070(03)00012-8

Tversky, A., Kahneman, D., Slovic, P., Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases Amos. *Science*, *185*(4157)arXiv arXiv:1011.1669v3, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uhl, M. W., Pedersen, M., & Malitius, O. (2015). What's in the News? Using News Sentiment Momentum for Tactical Asset Allocation. *The Journal of Portfolio Management*, *41*(2), 100–112. https://doi.org/10.3905/jpm.2015.41.2.100

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, *39*(2005), 165–210.

Zarowin, P. (1989). Short-Run Market Overreaction : Size And Seasonality Effects. *The Journal of Portfolio Management*, *15*(3), 26–29. https://doi.org/10.3905/jpm.1989.409209

Zhang, C. (2008). Defining , Modeling , and Measuring Investor Sentiment. *Working Paper*, (April), 0–42.

Zhang, W., & Skiena, S. (2010). Trading Strategies to Exploit Blog and News Sentiment. *Journal of Banking and Finance*, *34*(6)arXiv 1511.06464, 1288–1298. https://doi.org/10.1016/j.jbankfin.2009.11.025

Appendices

Appendix A

Henry's Finance Dictionary: List of positive and negative words

A.1 Henry's Finance Dictionary: List of positive words

above accomplish accomplished accomplishes accomplishing accomplishment accomplishments achieve achieved achievement achievements achieves achieving beat beating beats best better certain certainty definite deliver delivered delivering delivers encouraged encouraging enjoy enjoyed enjoying enjoys exceed exceeded exceeding exceeds excellent expand expanded expanding expands expansion good greater greatest grew grow growing grown grows growth high higher highest improve improved improvement improvements improves improving increase increased increases increasing larger largest leader leading more most opportunities opportunity pleased positive positives progress progressing record reward rewarded rewarding rewards rise risen rises rising rose solid strength strengthen strengthened strengthening strengthens strengths strong stronger strongest succeed succeeded succeeding succeeds success successes successful up

A.2 Henry's Finance Dictionary: List of negative words

below challenge challenged challenges challenging decline declined declines declining decrease decreased decreases decreasing depressed deteriorate deteriorated deteriorates deteriorating difficult difficulty disappoint disappointed disappointing disappointment disappoints down downturn drop dropped dropping drops fail failing fails failure fall fallen falling falls fell hurdle hurdles least less low lower lowest negative negatives obstacle obstacles penalties penalty risk risks risky shrink shrinking shrinks shrunk slump slumped slumping slumps smaller smallest threat threats uncertain uncertainty under unfavorable unsettled weak weaken weakened weakening weakens weakness weaknesses worse worsen worsening worsens worst

Appendix B

Nifty 50 and Sectoral Indices Factsheets

B.1 NIFTY50 Factsheet as on 27-11-2020

Table B.1: Sector Representation in NIFTY50 index

Sector	Sector Weight(%)
FINANCIAL SERVICES	39.28
IT	15.75
OIL & GAS	12.94
CONSUMER GOODS	11.16
AUTOMOBILE	5.58
PHARMA	3.54
METALS	2.46
CONSTRUCTION	2.45
CEMENT & CEMENT PRODUCTS	2.19
TELECOM	1.99
POWER	1.7
SERVICES	0.54
FERTILISERS & PESTICIDES	0.41

Table B.2: Top constituents by weightage in NIFTY50 index

Companys Name	Weight(%)
HDFC Bank Ltd.	11.21
Reliance Industries Ltd.	11.17
Housing Development Finance Corporation	7.23
Infosys Ltd.	7.21
ICICI Bank Ltd.	5.84
Tata Consultancy Services Ltd.	5.04
Kotak Mahindra Bank Ltd.	5
Hindustan Unilever Ltd.	3.42
ITC Ltd.	3.03
Axis Bank Ltd.	2.67

(Source: niftyindices.com)

B.2 NSE SECTORAL INDICES Factsheet as on 27-11-2020

Table B.3: Nifty Auto constituents

Companys Name	Weight(%)
Maruti Suzuki India Ltd	22.12
Mahindra & Mahindra Ltd	16.35
Bajaj Auto Ltd	9.77
Hero MotoCorp Ltd	9.55
Eicher Motors Ltd	8.35
Tata Motors Ltd	7.64
Motherson Sumi Systems Ltd	4.20
MRF Ltd	3.82
Balkrishna Industries Ltd	3.19
Ashok Leyland Ltd	3.14

Table B.4: Nifty Bank constituents

Companys Name	Weight(%)
HDFC Bank Ltd	28.4
ICICI Bank Ltd	18.63
Kotak Mahindra Bank Ltd	15.94
Axis Bank Ltd	14.47
State Bank of India	10.4
IndusInd Bank Ltd	5.4
Bandhan Bank Ltd	2.75
Federal Bank Ltd	1.39
RBL Bank Ltd	0.95
IDFC First Bank Ltd	0.86

Table B.5: Nifty FMCG constituents

Companys Name	Weight(%)
Hindustan Unilever Ltd	28.32
ITC Ltd	25.10
Nestle India Ltd	9.47
Britannia Industries Ltd	6.55
Tata Consumer Products Ltd	4.92
Dabur India Ltd	4.31
Godrej Consumer Products Ltd	4.40
Colgate Palmolive (India) Ltd	3.80
Jubilant Foodworks Ltd	2.92
Marico Ltd	2.90

Table B.6: Nifty Energy constituents

Companys Name	Weight(%)
Reliance Industries Ltd	26.8
Power Grid Corporation of India Ltd	13.6
NTPC Ltd	12.66
Adani Green Energy Ltd	12.26
Bharat Petroleum Corporation Ltd	8.26
Oil & Natural Gas Corporation Ltd	7.9
Indian Oil Corporation Ltd	5.93
GAIL (India) Ltd	5.24
Hindustan Petroleum Corporation Ltd	4.31
Tata Power Co. Ltd	3.03

Table B.7: Nifty IT constituents

Companys Name	Weight(%)
Infosys Ltd	26.12
Tata Consultancy Services Ltd	26.30
Wipro Ltd	9.83
Tech Mahindra Ltd	9.67
Info Edge (India) Ltd	9.46
HCL Technologies Ltd	9.00
Larsen & Toubro Infotech Ltd	3.88
MphasiS Ltd	3.13
MindTree Ltd	1.69
Coforge Ltd	1.18

Table B.8: Nifty Media constituents

Companys Name	Weight(%)
Zee Entertainment Enterprises Ltd	28.10
PVR Ltd	17.42
Sun TV Network Ltd	16.63
TV18 Broadcast Ltd	12.53
Inox Leisure Ltd	9.56
Dish TV India Ltd	5.78
TV Today Network Ltd	3.13
D. B. Corp Ltd	2.44
Jagran Prakashan Ltd	2.41
Hathway Cable & Datacom Ltd	2.10

Table B.9: Nifty Pharma constituents

Companys Name	Weight(%)
Dr. Reddy's Laboratories Ltd	19.64
Sun Pharmaceutical Industries Ltd	18.51
Divi's Laboratories Ltd	15.39
Cipla Ltd	12.69
Aurobindo Pharma Ltd	8.18
Lupin Ltd	7.18
Biocon Ltd	6.39
Torrent Pharmaceuticals Ltd	4.30
Alkem Laboratories Ltd	3.87
Cadila Healthcare Ltd	3.86

Table B.10: Nifty Private Bank constituents

Companys Name	Weight(%)
HDFC Bank Ltd	26.42
ICICI Bank Ltd	19.39
Kotak Mahindra Bank Ltd	16.59
Axis Bank Ltd	15.60
IndusInd Bank Ltd	9.45
Bandhan Bank Ltd	4.82
City Union Bank Ltd	2.67
Federal Bank Ltd	2.43
RBL Bank Ltd	1.66
IDFC First Bank Ltd	1.50

Table B.11: Nifty PSU Bank constituents

Companys Name	Weight(%)
State Bank of India	34.54
Bank of Baroda	17.47
Punjab National Bank	12.48
Canara Bank	10.40
Union Bank of India	6.47
Bank of India	5.33
Indian Bank 3	3.00
UCO Bank	2.41
Indian Overseas Bank	2.36
Central Bank of India	1.90

Table B.12: Nifty Realty constituents

Companys Name	Weight(%)
DLF Ltd	26.61
Godrej Properties Ltd	24.30
Phoenix Mills Ltd	12.65
Oberoi Realty Ltd	12.40
Prestige Estates Projects Ltd	7.64
Brigade Enterprises Ltd	5.45
Indiabulls Real Estate Ltd	4.13
Sunteck Realty Ltd	3.50
Sobha Ltd	2.58
Omaxe Ltd	0.70

Appendix C

Research Publications/Paper Presentations

C.1 Publication

Chari, S. G., Hegde Desai, P., & Borde, N. (2017). A review of literature on short term overreaction generated by news sentiment in stock market. Anushandhan, 7(1), 1221.

C.2 Paper Presentation

C.2.1 International Conference

1. Chari, S. G., Hegde Desai, P., & Borde, N. (August 12-13, 2016) Short term overreaction in Indian Stock Market - An Evidence from CNX Nifty 50 stocks". Paper presented at International Conference on “Financial Markets and Corporate Finance" organized by Department of Management Studies, IIT Madras. Paper id: ICFMCF13022

2. Chari, S. G., Hegde Desai, P., & Borde, N. (March 28-29, 2016) “Does News sentiment generate overreaction in stock markets". Paper presented at the International Conference on “Emergence of India as a Global Power- Challenges and opportunities" organized by S. S. Dempo College of Commerce and Economics, Goa.