Can the Sentiment in Print Media during National Elections Influence Stock Market Dynamics? - A Case-Study on 2019 Indian Elections

Manvitha Kola

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Prof. Khurshid Ahmad

August 2021

Declaration

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

Permission to Lend and/or Copy

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

Can the Sentiment in Print Media during National Elections Influence Stock Market Dynamics? - A Case-Study on 2019 Indian Elections

Manvitha Kola, Master of Science in Computer Science

University of Dublin, Trinity College, 2021

Supervisor: Prof. Khurshid Ahmad

This research work is an attempt to statistically study the variance in the stock market returns based on sentiment expressed in print media. Stock markets are very complex interrelated systems that are influenced by various factors including investor sentiment and information sources. This thesis employs a statistical approach to take a closer look at that relationship, if exists. The content in the print media articles is analyzed to extract the sentiment as a time series variable. This is used as an explanatory variable to develop various statistical models to determine the impact of media sentiment on the changes in stock market returns. This is a case study of a political event, i.e. the 2019 Indian General Elections. The print media coverage during a political event may involve various biases to distort political sentiments and opinions of public. Therefore, the Indian news articles containing election news have been explored to investigate if there is any statistically significant impact of the news articles on the stock prices based on the repeated political or affect terms during the general elections. These repeated political and affect features signify positive/negative sentiment thereby endorsing or criticizing certain categories. Rocksteady, an in-house affect analysis system developed at Trinity College Dublin was used for this analysis. The news articles published in the major newspaper publications of the Indian media between 1st Jan 2018 and 31st Dec 2020 - during the event of National general elections in 2019 and 2020 are analyzed and the statistically significant results are published. It has been observed that the media coverage for the winning party was higher when compared to that of the opposition and there is a statistical evidence that various political variables have a small but significant effect on the stock market dynamics.

Acknowledgments

Firstly, I would like to express my sincere gratitude to Professor Khurshid Ahmad, my supervisor, for his continuous support and constant motivation to carry out this research and make this dissertation a reality. Without his guidance, this project would not have been possible. I am very grateful for the opportunity to learn from him and I would like to express my sincere thanks for his endless encouragement and amazing guidance.

I would like to thank Sachin, Bhavana and all my dearest friends for their continuous support and encouragement in times of pandemic and throughout the process of writing this thesis.

Lastly, I would like to express my most heartfelt thanks to all my family, Mom, Dad and my brother for being a great support throughout my life. Without you, any of this wouldn't be possible.

Manvitha Kola

University of Dublin, Trinity College
August 2021

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As leading political economist Edward Tufte stated, "When you think elections, think economics" [William and Tufte (1979)]. The term "Election", as defined by Wikipedia, is a formal group decision-making process by which a population chooses an individual or multiple individuals to hold public office. Studies that date as far back as the 1930's support the idea that elections and economy are somehow correlated [Tibbitts (1931)]. In an emerging market like India, empirical evidence suggests that there exists a positive correlation between economic governance and election prospects, that is, economy impacts the electoral outcome [Gupta and Panagariya (2014), Vaishnav and Swanson (2015),Sharma and Swenden (2020)]. This study is a research conducted to understand the influence of sentiment expressed in Indian English print media during 2019 general elections on the stock market dynamics, if there is any.

"Hence, in order to have anything like a complete theory of human rationality, we have to understand what role emotion plays in it" - Herbert Simon, 1983

Nobel Laureate Herbert Simon, in his book "Reason in Human Affairs" draws attention to the importance of identifying emotion as a significant factor in explaining human rationality [Meehan (1983)]. There has been increasingly growing research on the effect of emotion and sentiment on judgement and decision making in a wide range of fields from Marketing [Achar et al. (2016)] to Neuroscience [Phelps et al. (2014)]. In the field of economics, the role of emotion in decision making attracted very little attention until recently, despite previously existing influential works [Loewenstein and Lerner (2003)] .In the recent times, there is a growing research in the domain of finance to analyse the text in news as a proxy for investor sentiment [Liu and McConnell (2013)].

In the field of cognitive psychology, many researchers believe that emotions influence and drive an individual's judgment and decision making process [Josephs (2005), Frijda (1988), Keltner and Lerner (2010)]. Emotions are defined as subjective feelings and thoughts which are closely related to sentiment and can have various intensities [Liu (2012)]. Though emotion is the widely used term to describe a strong feeling, a broader term "affect" is used in psychology, which encompasses feelings, emotions and moods. Affect is defined as an experience of feeling or emotion. Affect can be both positive and negative. Words that contain any kind of emotion or affect are termed as "affect words". They are majorly classified along three dimensions: valence, arousal, and strength. As words such as "happy", "delighted", "excited" create a positive affect, they are classified as positive affect words and words such as "sad", "miserable", "depressed" trigger negative sentiment and are thus classified as negative affect words.

Sentiment analysis is one of the most active and widely researched areas in NLP in the recent times. According to [Liu (2012)], Sentiment analysis can be defined as the computational study that analyzes an individual's opinions, sentiments and emotions from text. There are many different names and tasks under sentiment analysis which include opinion mining, opinion extraction, subjectivity analysis, affect analysis, etc. [**?**]. It is a subset of Natural Language Processing(NLP) which focuses on analyzing, extracting and quantifying affect information from text. It uses affect words to extract the emotions and underlying sentiment associated with the text. There are several approaches to analyze sentiment in text. Lexicon based sentiment analysis is one of the major approaches to analyse sentiment using the concept of bag-of-words. In this approach, the text is represented as a bag of words and a sentiment dictionary is used to calculate the affect words within the text. An aggregation function such as sum or average is used to calculate the overall sentiment in the text.

This chapter introduces the notion of sentiment in Media, Politics and Stock Market and briefly describes the relation between them along with providing a background to support the case-study. It is followed by the Research objectives and concludes with a brief outline on the thesis layout.

1.1 Background

1.1.1 Role of Media during General Elections

In democratic states, the media has a significant role in influencing public opinion [Yao et al. (2012), Ullah and Khan (2020)]. Especially during political events such as elections, the media acts as a campaign platform and a forum for debate and discussion, becoming one of the primary players in influencing the electoral outcomes [Craig (2004)]. Print media is one of the oldest mass media platforms with great diversity in terms of ownership and content. Though the other mass media platforms such as television and social media have a significantly larger audience, newspapers still play a major role in creating various degrees of impact in public opinion based on various economic, political and geographical factors. Along with the responsibility of creating awareness of issues in the country and keeping the citizens informed of the current events, the media poses a risk of introducing various biases by manipulating text to meet individual interests.

While the print media is witnessing a global decline, India is one of the few countries in the world where the print media is not only dominant, but also experiencing a significant surge in newspaper consumption in recent times. As per the data released by Indian Readership Survey (IRS), the number of newspaper readers have grown from 407 million in 2017 to 425 million by the end of March 2019. Though English newspapers have a very small reader base, when compared to Hindi and other regional dailies, there is still a surge of 10.7% growth in English newspaper readership from 28 million to 31 million readers between 2017 and 2019 as per IRS[(RNI)].

1.1.2 News as Political sentiment Proxy

Media is a means of mass communication. The main purpose of any media is to keep the society informed and transfer knowledge. There are various kinds of media which includes traditional media for mass communication such as the Print media, Television, Radio and digital media such as social media and online news. Though with the advent of the digital era, online media is witnessing a major upsurge in recent times, online disinformation and fake news still remain a key issue to be tackled. One major advantage of Print media over other media such as online or social media is the in-depth fact-checked reporting and analysis provided in the newspaper articles.

Media plays a major role in the development of any country, especially in times of political events. It acts as a mediator between politics and the general public. Though

the media is an established source of political information, it is still subjected to various biases which influences voter's perceptions thereby manipulating their democratic opinion-forming [Mutz (1992)]. The current day media holds enough power to manipulate the voter perception through its political narratives.

During the times of political events such as the General elections, Political parties leverage mass media to influence the electoral outcome. Lot of research has been conducted to understand how media can influence the voter's opinion[Nofsinger (2001), Eberl et al. (2017), Craig (2004), Farrell (1987), Petrocik et al. (2003), Rogers et al. (2006), Ahmad et al. (2011)]. It has also been observed that frequent exposure to certain media also has an impact on effecting the electoral outcome[Gerber et al. (2009)].

As media is considered to be one of the key factors in influencing voter political sentiment, news articles in the print media are considered as a proxy for political sentiment for this case study.

1.1.3 News as Stock Market sentiment proxy

Stock market dynamics are forces that impact stock market prices and consumer behaviors. The stock market prices are generally determined by supply and demand of shares in the market. However, there are many other factors such as the economy, investor sentiment, etc. that influence the stock prices.

Media can play a major role in explaining the fluctuations in the stock market. Narratives in the media, especially which are of human interest and emotion , can be considered to understand the fluctuations in the economy [Shiller (2015), Shiller (2017)]. Factors such as volume of articles and the number of words in each media article are also said to impact the stock market prices. Research suggests that the frequency of positive and negative words in the print media article content can be used to measure the sentiment in the newspaper reporting [Ahmad et al. (2011)]. Lot of research has been conducted in recent times to understand the role of sentiment on stock market prices. The stock market dynamics are highly correlated with the investor sentiment and has a significant and regular effect on the stock market [Baker and Wurgler (2007)]. Tetlock (2007) studied the role of print media in predicting daily stock returns and found that the negative sentiment expressed in the wall street journal market column influences the market behavior. Hence, the sentiment expressed in the media articles can have an influence on the overall mood of the stock market.

In this case study, the tone of the major Indian newspaper articles which is measured as the frequency of number of positive and negative words used in each newspaper article is used as a proxy for investor sentiment. The daily measure of sentiment or negative tone for a period of 3 years from 2018 to 2020 is considered as a sentiment proxy for this research.

1.2 Research Objectives

This research aims to study the stock market dynamics during the 2019 Indian General Elections by examining the sentiment expressed in the Print media as well as various other political features and its impact on the movement in the Indian stock market, if any, for a time span of three years from 2013 to 2015. The two main objectives of this research are to

1. Investigate the relationship, if exists, between the negative sentiment in print media and the stock market returns during a political event
2. Statistically assess the obtained vector auto regression model results to investigate if there exists any statistical significance of negative sentiment and other political features on stock market dynamics.

1.3 Layout of Thesis

Chapter 2 describes the state of current research of various techniques employed in this study. Chapter 3 explores various research methods used in the research and the theoretical background of different techniques and methods employed. Chapter 4 is the Case study and results chapter, where the case study is explained and the results of the thesis are presented. Final chapter is the Conclusion which summarizes the work, and describes the challenges encountered and the possible future work based on the current case study.

Chapter 2

Motivation and Literature Review

2.1 Motivation

It is intuitive to assume that the media has an influence on human perceptions and behaviors. News in print media is one of the major information sources and studying the impact of media in influencing opinions and decision making has become a topic of interest in various domains including politics and economics. With the rise in social media and the increased availability of huge volumes of user generated data, a lot of research is focused on understanding the overall user sentiment expressed in social media to analyze and influence the user behavior. There is increasing research in understanding how to leverage media as a political campaigning tool to influence voter perception and behavior during elections. In the stock markets, uncertainty and market sentiment influence the stock market prices, volatility and dynamics. The investor's mood/belief is one of the key stock market predictors, and can be influenced by the flow of information in various media channels by changing the beliefs and effecting their sentiment.

Sentiment analysis is a growing research area in recent times. With the rise in social media and the increased availability of huge volumes of user generated data, a lot of current research is focused on understanding the overall user sentiment expressed in social media to analyze and influence the user behavior. There is also increasing research to understand how the media can be leveraged as a political campaigning tool to influence voter perception and behavior. It is natural to assume that media has an effect on the stock market price movements.

News and media are one of the major sources of information and they are widely consumed by the general public which include market investors. Research evidence proves that the

media tends to influence and distort the opinions and sentiments of the general public [Ahmad et al. (2011)]. The term 'Investor sentiment' in the domain of finance indicates the general mood among the investors which is said to be influenced by the news and media.

The news in the media contain a great wealth of information which can be leveraged to model the financial model behaviors. There has been increasing research in the finance domain to analyze the news content as an additional variable to predict the financial market dynamics. Research studies find that the political sentiment expressed in news media has an influence on the financial markets [Kelly and Ahmad (2018), Chan and Lakonishok (1993), Barber and Odean (2008), Fang and Peress (2009), Engelberg and Parsons (2011)]. Thus, sentiment expressed in media is an ideal explanatory variable that influences investor behavior and thereby stock market dynamics. This case study is an attempt to understand and assess the relationship between sentiment in media and the changing stock market behavior during 2019 Indian general elections.

2.2 Literature Review

2.2.1 Sentiment Analysis

According to [Liu and McConnell (2013)], "Sentiment Analysis is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language". Sentiment analysis mainly studies opinions which express or imply positive or negative sentiments. In general, Sentiment and Opinions are one of the key factors in influencing human behavior [Liu and McConnell (2013)]. As per the definition in Oxford English Dictionary, Sentiment analysis is "the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral."

Words and phrases that convey positive or negative sentiments are instrumental for sentiment analysis [Liu and McConnell (2013)]. These words are also known as "sentiment words", "opinion words" or "affect words" in the research literature. Few examples of positive affect words include beautiful, wonderful, and amazing whereas examples of negative sentiment words include bad, awful, and poor. As most dictionaries list synonyms and antonyms for each word, using a dictionary to compile sentiment words is an obvious approach [Miller (1995)]. Kamps et al. (2004) had suggested a much more sophisticated

technique to calculate sentiment which used a WordNet distance-based approach to calculate the sentiment orientation of a given adjective. Lexicon-based method for sentiment analysis is an unsupervised approach to compute the sentiment in a text document using a dictionary of sentiment words along with their associated polarity and strength [Taboada et al. (2011)].

In this case study, a lexicon-based approach has been used to extract sentiment from News article corpus. Using this approach, affect dictionaries of pre-tagged words associated with the affect category are used to calculate the frequency of those words in the given text to determine the affect value. The General Inquirer lexicon [Grieb (1968)] is one of the several general-purpose sentiment/affect dictionaries that is publicly available to extract sentiment in text using a dictionary-based approach, and has been used in this case study. An additional case-specific dictionary has been created for the purpose of this research in relation with the collected corpus that has been appended to the previously discussed General Inquirer dictionary. Although a dictionary based approach is quick and easy to apply, one disadvantage of using this approach is that the sentiment polarity of words analyzed using this technique are general, i.e. they are domain and context independent.

2.2.2 Sentiment in media during Elections

"Elections define democracy while the media enlightens and sustains it"
- Peter Forau, Pacific Islands Forum, 2006

Conventional media such as newspapers and television as well as the social media spread information about significant political events. The print media crafts the political discourse by choosing which political party to cover and how much to cover and reflect political biases in their coverage, which alone can influence voter perceptions. Various major political events such as the German Federal Elections in 2009 US Presidential Elections (2008) , the General Elections in Belgium (2011) have been studied to understand the influence of bias in the social media on the voter opinion [Jürgens et al. (2011), Mullen and Malouf (2006), Junqué de Fortuny et al. (2012)]

In his research, Craig (2004) studies the role of the news media in representing politics and shows how politicians can leverage the news media. Various biases such as the coverage, association, and subjective language bias in the news media which frames information so as to influence the reader's opinions [Sales et al. (2019), DellaVigna and Kaplan (2007)] investigated if media bias affects voting and found that introducing Fox news into

the cable markets significantly affected voter turnout and the Republican vote share in the Senate.

The frequency of named entities in the media may have an effect on the voter judgements.Ahmad et al. (2011) provides evidence of coverage bias in 2011 Irish elections as Male candidates are cited 8 times more than the female candidates. It is also observed that the citation of political parties and leaders had a strong correlation with the electoral outcome [Ahmad et al. (2011)].

Ahmed et al. (2017) investigated the use of social media platform, Twitter as an e-campaign tool during the 2014 Indian general elections. It is observed that the success of a winning party is highly correlated with their twitter usage during election campaigns to engage young voters. Bharathiya Janatha Party(BJP) which was then an opposition party, had drastically improved their chances of winning by engaging the youth voter base by leveraging the social media interactivity.Cameron et al. (2016) adopted a forecasting model to predict the electoral outcome of the 2011 New Zealand general election using social media data as an indicator. Their research suggests a statistically significant relationship between social media usage and election results.

2.2.3 Effect of sentiment in Stock Market

Research finds that stock market returns are influenced by the investor sentiment [Brown and Cliff (2005), Baker et al. (2012), Ho and Hung (2008)]. Hence, it can be assumed that if investor sentiment is influenced by various political variables, stock price returns might also be influenced by the political features. Market sentiment has a major influence on the stock market dynamics. Here we consider the sentiment in the newspaper media as a proxy for 'market sentiment'. The investors in stock markets generally rely on information that is publicly available and easy to access.

Several research studies have investigated the trading behavior of investors and the stock market price impact [Chan and Lakonishok (1993) , Keim and Madhavan (1997)] . Tetlock (2007) conducted a systematic study of the relationship between media content and stock market activity to predict movements in stock market by analyzing the sentiment in the Wall Street Journal columns. He employed vector auto regression to predict Dow Jones price returns using negative Sentiment score. He states that "who read the Dow Jones News wires can devise profitable trading strategies by calculating the daily value of pessimism/negative sentiment to forecast future returns."

Yang et al. (2019) explore the relationship between media coverage and stock returns in China using major Chinese newspaper media reports. They compared the one-year period and noticed that the firms with high news media coverage in the current month have better stock price returns in the next few months. They observed that the news media has a significant impact and a positive correlation with price returns in the Chinese stock market which mostly consists of individual investors or immature investors.

It has also been observed that the volume of news articles also has a major impact on the investor sentiment and stock market returns. Research was carried out to analyze the sentiment based on the volume of news items on a financial topic in a given period of time [Antweiler and Frank (2006)]. Though the news article's volume does not impact the sentiment directly, multiple mentions of a term in the news indicates a major event or activity around it, which is an indirect measure of sentiment.

The visibility of public information plays a key role in behaviors of the stock market investors [Nofsinger (2001)]. The consumer sentiment and their perceptions of the economy is affected by the news media through three main channels which include latest economic data and the opinions of professionals, tone and volume of the economic reporting and volume of news about the economy [Nofsinger (2001)]. It is also found that the expectations of consumers about the economy are updated more frequently when there is high news coverage especially during and after events like recessions.

Mitchell and Mulherin (1994) studied the interaction between the count of news articles reported by Dow Jones & Company and its impact on securities market activity including market returns and trade volume. Fang and Peress (2009) used the number of the newspaper as the measure of media coverage to explore the effect of media coverage on stock returns and found that there is a higher return for the stocks with no media coverage than the stocks with high media coverage.

Economists had researched various topics in media including weather, sports, pollution and their effect on stock market dynamics [Symeonidis et al. (2010), Edmans et al. (2007), Levy and Yagil (2011)]. Symeonidis et al. (2010) investigates the relationship between volatility in the stock market returns and investor mood-proxies related to the weather and the environment (cloudiness, temperature, precipitation, night time length). Edmans et al. (2007) investigates the changes in stock market prices to sudden changes in investor mood based on the psychology research evidence of a strong correlation between the soc-

cer game results and sentiment. Based on Health related research that air pollution has negative mood effects which lead to increased risk aversion, Levy and Yagil (2011) investigates whether the mood effects caused due to air pollution can have any impact on stock market returns.

Over the last few years, significant research was conducted to analyze the sentiment in social media content such as blog content and in particular large-scale Twitter feeds. Tweets are individual user posts with a limitation of 140 characters. Millions of tweets at any point in time are considered as proxy for public sentiment. The social media sentiment is claimed to be a major factor in predicting the stock market behavior [Bollen et al. (2010), Mittermayer and Knolmayer (2006)]

Santa-Clara and Valkanov (2003) evaluate the impact of various political features on stock market returns and volatility. In their study, the financial variables such as the stock price returns are regressed against the investor sentiment and a dynamic conditional correlation model was employed to verify the impact of political features on investor sentiment and stock market returns

Bollen et al. (2010) studies the extent of public mood, as expressed within a large corpus of tweets, correlates with stock prices. Yu et al. (2013) conducted a research to compare and contrast the relative importance of social media and conventional media along with their inter-relatedness on short term firm stock market performances. Social media includes individual blog posts, twitter, etc whereas conventional media consists of major newspapers, business magazines. It has been observed that overall social media correlates strongly with stock performance than conventional media. It has also been observed that social media is interrelated to conventional media to influence the stock market.

However, in recent times, there are growing concerns in the reliability of public opinion in social media. fake news is on the rise. Fake News is defined as a fabricated article with the intention to mislead, usually for profit [Harjule et al. (2020)]. According to the recent study conducted by Tayal and Bharathi (2021) "It was found that most of the social media posts about COVID-19 were not so reliable and trustworthy. However, social media posts which are shared by competent persons such as doctors, medical practitioners etc. were trusted by the people".

There is an ongoing research by Ante (2021) which shows that there is a significant impact of social media activity of highly influential and well-known individuals on cryp-

tocurrencies market prices. For example, On January 29, 2021, Elon Musk, at that point in time, was the richest person in the world (Klebnikov, 2021). The moment he changed his twitter bio to #bitcoin, with in a matter of hours, the market capitalization of bitcoin increased by $111 billion from about $32,000 to over $38,000 within few hours which indicates that a single tweet has a potential to increase of $111 billion in Bitcoin's market capitalization and alternatively a tweet can also drop out the market value of similar magnitudes indicating the significant impact of fluctuations in public sentiment based on social media content. Garrett (2019) studied the contribution of social media to the political misconceptions during 2012 and 2016 U.S. Presidential elections.

Chapter 3

Research Methods

This chapter describes the methods and techniques employed during this research. This chapter is divided into 4 major sections namely Data acquisition, pre-processing and transformation, Affect Analysis, Data exploration and visualization, Statistical Time series Analysis. Implementation of the case study is described in the final section of the chapter. Each section contains a detailed explanation of the methods and techniques used in carrying out the research to achieve the research objectives

3.1 Data Pipeline

This section explains the 3 major steps implemented in creating the data pipeline for this research which include Data acquisition, Data pre-processing and Data transformation. Data acquisition can be defined as the process of sampling real world data and converting the collected data into numeric values that can be processed by a computer. Data preprocessing is a technique which is used to transform the collected raw data into the required format to perform the computation. Data transformation is a process of cleaning, formatting, filtering, computing and integrating the data.

3.1.1 Data Acquisition

Data acquisition is the first step in the data pipeline. To analyze the sentiment in the print media, news article data is collected. LexisNexis News & Business digital database was used to collect the news articles published in major Indian English newspapers during the election timeline. LexisNexis is an online research database which maintains the corpus of worldwide major newspapers. It allows users to filter news based on keywords, date ranges, publication type and name as well as many other useful features.

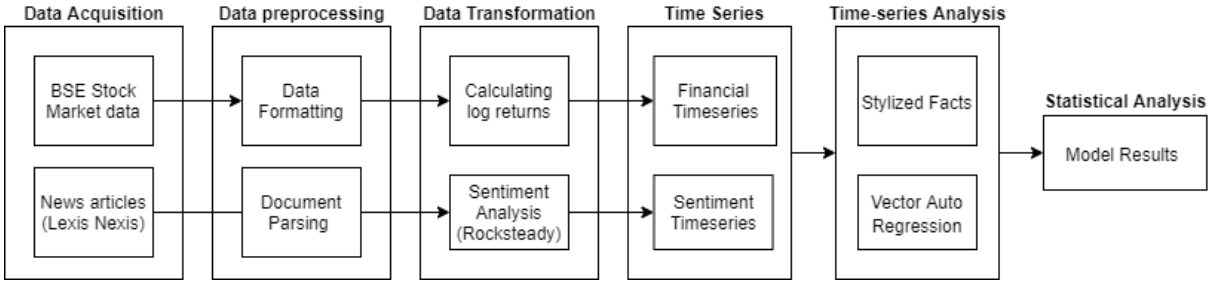

Figure 3.1: Data Flow diagram: News articles (LexisNexis) and BSE SENSEX Data was acquired which is then pre-processed into the required format. The Sensex data is transformed into log returns and sentiment is extracted from the newspaper corpus to generate a VAR model results which are then verified for statistical significance

A key-based data selection criterion has been implemented to select only the news articles from the LexisNexis database that mention the term Elections(s) 3 or more times anywhere within the news article. The term 'election*' was given in the query search, such that, the articles containing the term 'election' are retrieved. The data range field in the search has limited the timeline from Jan 1st 2018 to Dec 31st 2020 to filter the news articles accordingly. LexisNexis also provides an option to automatically remove the duplicate news articles. LexisNexis provides the option to download the news articles into a single file containing 500 news items in one single batch. The collected data is then organized in a consistent folder structure by Newspaper publication type for each quarter in the collected date range.

Five major Indian English newspaper publications as issued by the Audit Bureau of Circulation are considered for the purpose of this research. The timeline of the study spans from 2018 to 2020 and the collected news article corpus consists of 50381 articles. These articles are organized on a quarterly basis for a span of 3 years, segregated by the publication. The newspapers chosen for this case study are as follows:

Table 3.1: Table showing Publication name, Founded year, Readership of the publication in millions as per Indian Readership Survey (IRS 2017) and number of articles collected from LexisNexis

Publication Name	**Year Found**	**Readership**	**No. of Articles**
The Times of India	1838	13.047	11311
The Economic Times	1961	3.103	8697
The Hindustan Times	1924	6.847	11062
The Indian Express	1932	1.599	10872
The Telegraph	1982	1.558	8439

The financial stock market SENSEX data was downloaded from the archives of the official of Bombay Stock Exchange website. The collected Sensex data is available on a daily basis with missing values for the weekends and other holidays.

3.1.2 Data Pre-preprocessing

Data pre-processing is the technique of transforming the collected raw data into the required format to perform the analysis. The newspaper articles collected from the Lexis-Nexis database in the previous data acquisition step are in the rich text file (RTF) format. These articles are to be converted into a rocksteady readable format such that the affect analysis calculations can be performed on the text corpus. A python based document parser is used to convert the downloaded news media articles in the RTF format to a rocksteady input format. The news media articles also contain a lot of metadata information about the articles including Article Load Date, Language, Publication Type, Article Title, By-line, section, length of the article which are extracted using the document parser. The news articles are organized on a daily time-series basis and are parsed accordingly to calculate the affect/sentiment for each day over the given period of 3 years. The data types and formats of the collected data were modified in order to facilitate calculations and analysis. This step includes converting both the sources of data, i.e. news article corpus and the SENSEX data into time-series format.

3.1.3 Data Transformation

Data transformation is the process of cleaning, formatting, filtering, computing and integrating the data. In financial market analysis, a commonly used measure for rate of return is the logarithm of the ratio (Log Ratio) of successive prices to get a more stationary financial time series data [Ultsch (2008)]. Returns are defined by changes in the logarithmic prices [Taylor (2007)]. So the closing prices of the stock return are transformed into a log ratio time-series.

Let p_t be the closing price stock index at time t, p_{t-1} be the closing index of the previous day, the return r_t over the time period from t to $t-1$ is given as,

$r_t = log(\frac{p_t}{p_{t-1}})$

$r_t > 0, if, p_t > p_{t-1}$, indicates increase in price, i.e. profit

$r_t < 0, if, p_t < p_{t-1}$, indicates decrease in price i.e. loss

$r_t = 0, if, p_t = p_{t-1}$, indicates no change in price

where returns can be negative, positive or zero depending on the previous day's closing index.

The closing prices from the collected BSE Sensex data is transformed into returns time-series dataset to analyze the stock market. Data integration is also a key task in data transformation. Both the time series data sources (i.e. sentiment extracted from the news corpus on a daily basis and the log returns of the stock prices for each day) are integrated to be fed into the GRETL software for regression analysis. Data cleaning and filtering is performed as the data from two sources has to be merged and the time series data of BSE Sensex is not continuous due to holidays. Thus, only the non-holiday data is considered for this time series analysis.

3.2 Sentiment Analysis

This case study implements a natural language processing (NLP) technique which is called the General Inquirer method developed by [Grieb (1968)]. This method uses the concept of Bag of Words which converts the text corpus into words/tokens and calculates their frequencies removing the context. The General Inquirer (GI) method takes the bag of words as input, compares it with the dictionary that contains terms under different categories and returns the number of times a term in the dictionary appears in the text.

Rocksteady is the sentiment analysis tool used for this case study, which is an affect analysis system developed in-house at Trinity College, University of Dublin (Ahmad et al., 2011). Rocksteady calculates the affect score of a text corpus and converts it into a time series. Rocksteady takes news articles corpus data as input and outputs a sentiment time series, which is used to develop the financial models in this research.

Rocksteady uses a combination of various general purpose affect dictionaries such as the General Inquirer dictionary along with an optional custom domain specific dictionary to calculate the sentiment score of the collected text corpus. The dictionary includes various affect and domain specific features such as negative, positive sentiment or election specific categories. It uses a frequency count approach that calculates the frequency of terms for each category in the integrated dictionary on a daily basis.

Negative sentiment is considered for this case study as it is observed from the previous research that the negative sentiment in the media is one of the major factors that influences investor sentiment [Tetlock (2007)]. Rocksteady provides 3 metrics to evaluate the sentiment in the text corpus which includes the total count/frequency of the negative affect words in the text corpus, percentage of negative word count and the standardized z-score values for the negative terms. The z-score metric is chosen for this case study as

it has been observed from the previous research that the residuals for stock price returns show the greatest movement when the z-score of negative terms are either greater than 1 or less than -1. These z-scores indicate the number of standard deviations away from the mean.

3.3 Data Exploration and Visualization

Data visualization is one of the most powerful processes that represents the data graphically to understand the patterns and hidden insights. It is the most efficient way of understanding the data and interpreting the results. As a part of the case-study, several visualizations are developed using 2 major data visualization tools: Tableau and Python

3.3.1 Tableau

Tableau is one of the most popular data visualization tools used to create various graphs and reports quickly for analysis. It supports multiple ad-hoc visualizations and can extract data in various formats to develop visual analytics. When compared to traditional tools like Microsoft Excel, Tableau is a more sophisticated tool to develop visualizations. It uses a drag-and-drop feature to convert data into graphs which can be easily understood and interpreted. Tableau also supports interactive visualizations to explore data using a wide range of visualizations.

3.3.2 Python

Data visualization is one of the major features supported by python. The two main libraries in python that support data visualization are Matplotlib and Seaborn. These libraries support a lot of features to develop customizable and interactive graphs and plots. The Matplotlib library is used for two-dimensional array plots which uses the Numpy library. It supports a wide range of static and interactive visualizations and a range of plots like line, bar, scatter, histogram, etc. for analysis. Seaborn is another important python library used for data visualization which is integrated with Pandas library. It is developed on top of Matplotlib library to support a wide range of visualizations.

3.4 Statistical Time series Analysis

This case study involves statistical analysis of sentiment and financial time series data. Both the time series data sets are combined on a daily basis for a period of 3 years (2018-2020). Two major statistical analyses are performed in this case study. Stylized facts are used to describe and summarize the time series data. Vector autoregression is a statistical model which is used to capture the relationship between multiple quantities as they change over time. In this case study, we use vector autoregression to determine the relationship between sentiment and stock price returns.

3.4.1 Stylized Facts

Stylized facts are used to describe continuous time-series data. They contain information about summary statistics along with the shape and central tendencies of the distribution. The key stylized facts are discussed below:
1. Mean (μ): Statistical mean is a measure of central tendency of a distribution. For a variable X with n observations $X_1, X_2, ..X_n$, the mean is given by,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. Median: Median is a statistical measure of central tendency of a data which is calculated by ordering the data and selecting the middle value. It is an efficient metric as it is not skewed by the outliers. It indicates where the data is concentrated.

3. Standard deviation: Standard deviation is a statistical measure of the spread of the data in the distribution which is obtained by calculating the sum of squared differences of each observation from the mean.For a variable X with n observations $X_1, X_2, ..X_n$, the standard deviation is given by,

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \mu)^2$$

4. Range: Range is calculated as the difference between the maximum and minimum value in the data, which indicates the spread of data in the distribution.

5. Skewness: Skewness is a statistical measure that characterizes the degree of asymmetry of a distribution around its mean.For univariate data $Y_1, Y_2, ..Y_n$, the Fisher-Pearson coefficient of skewness formula for skewness is given as:

$$Skewness = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\mu)^3}{\sigma^3}$$

where μ is the mean, σ is the standard deviation, and n is the number of data points.

6. Kurtosis: Kurtosis is a statistical measure of the relative peak or flatness of a distribution as compared to the symmetric normal distribution.For univariate data $Y_1, Y_2, ..Y_n$, kurtosis is given as:

$$Kurtosis = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\mu)^4}{\sigma^4}$$

where μ is the mean, σ is the standard deviation, and n is the number of data points.

The skewness for a normal distribution is zero, i.e. any symmetric data has a skewness of nearly zero. Negative values for the skewness indicate that the data distribution is left skewed which means a longer left tail when compared to the right tail. Positive values for the skewness indicate that the data distribution is right skewed which means that the right tail is longer when compared to the left tail. Significant skewness and kurtosis clearly indicate that data are not normal.

3.4.2 Linear Regression

To understand the VAR model, let us first look at linear regression. In statistics, linear regression is defined as a linear approach to model the relationship between a dependent variable and one or more independent variables.

Let us assume two variables X and Y where X is the independent variable and Y is the dependent variable. The relationship between the variables X and Y is given by the following equation:

$$Y = \alpha X + c + \epsilon$$

where α is the coefficient, c is the constant and ϵ is the error term.

The above linear regression equation models the statistical relationship between two variables X and Y, where Y depends on the independent variable X. α is known as the regression coefficient which means that if X differs by one-unit Y will differ by α units on average. The constant term c represents the average change in the dependent variable and the error ϵ is a random variable which adds "noise" to the relationship.

3.4.3 AutoRegression

Auto regression is a statistical time series modelling technique which uses observations from previous time steps as input to a regression equation to predict the value at the next time step. As the model uses the observations of the same input variable at previous time steps, it is referred to as auto regression The predicted value of variable X at a time step t+1 is given by:

$$X_{t+1} = b_0 + b_1 X_{t-1} + b_2 X_{t-2}$$

Where b_0 is the constant and b_1 and b_2 are the regression coefficients of the variable X at times t-1 and t-2 respectively.

3.4.4 Vector auto regression

A common assumption made in economic analysis is to study the effect of past values of an economic variable to model the current value of that variable. Vector autoregression is one of the most commonly used statistical frameworks in macro econometric studies to forecast economic time series proposed by Sims. VAR model suggests that a variable is dependent on the lagged values of itself (Sims, 1980). For a univariate analysis, this can be expressed as:

$$r_t = f(r_{t-1}, r_{t-2}, r_{t-3}...)$$

Let us suppose the return value is denoted by r and rt denotes the return value at time t. The returns on the previous days are denoted by r_{t-1} (one-day lag), r_{t-2} (two-day lag) and so on. Then, by Sim's VAR model, the value of rt is dependent on the previous values of r. The number of previous variables to be considered for the model are known as 'lags' (). The VAR model equation of univariate time series analysis for a 5-day lag is as follows.

$$r_t = C + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \alpha_3 r_{t-3} + \alpha_4 r_{t-4} + \alpha_5 r_{t-5} + \epsilon_t$$

where, $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are the coefficients, ϵ is the residual term and C is the constant. The above VAR model equation can also be expressed as follows:

$$r_t = C + \alpha_t r_t L_{\tau=5} + \epsilon_t$$

where α_t is the coefficient, r_t is the return value at time t and ϵ_t is the error term at time t for $L_{\tau=5}$ 5 lag values

Sims VAR model extends this to a n-variable linear model in which each variable can be explained as a function of linear combination of its own lagged values, plus current and past values of the remaining n-1 variables.Unlike the auto regression models that use only their past values to model a variable, the VAR model considers other variables along with the interaction between them to model the dependent variable. Thus, the inter-variable dependencies and relationships can be interpreted using a VAR model. The model results include the regression coefficients using the direction of correlation and the magnitude (multiplying factor) can be determined. An additional statistical test is performed to verify if the model results obtained are statistically significant.

3.5 Tools and Softwares

This section describes the tools and softwares used for the implementation of this case-study.

3.5.1 LexisNexis

Lexis Nexis News Business is an online digital database research tool for news, companies, markets insights and risk assessment. It maintains a collection of worldwide media corpus. LexisNexis provides computer based services for legal research and risk management. It allows users to filter news articles based on various categories such as keywords, dates, publications, industry, etc. The articles can be downloaded in 3 major formats which include PDF, MS Word (docx) and Rich Text Format (rtf). LexisNexis also provides various formatting options before downloading the documents. The meta data of the news articles is also included in the documents, which provides additional details regarding the documents such as the publication type, subject, industry, language, location, etc.

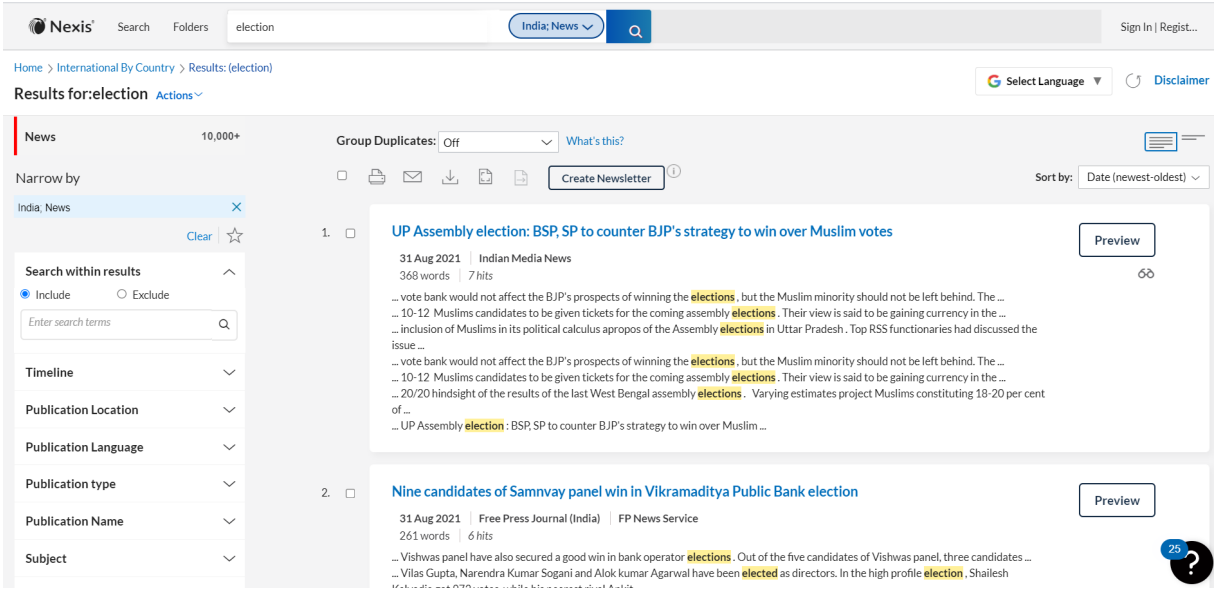

Figure 3.2: A Snapshot of Lexis Nexis User Interface

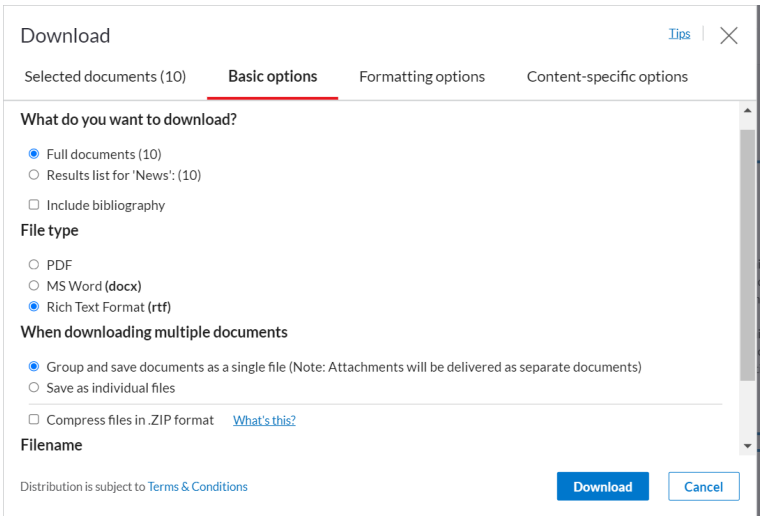

Figure 3.3: A Snapshot of LexisNexis UI window which contains basic options to download news corpus

3.5.2 Rocksteady

Rocksteadyis an in-house affect analysis tool, developed at Trinity College Dublin. Ituses a dictionary based approach to calculate the affect/sentiment in a textcorpus. Rocksteady uses the General Inquirer dictionary which is a list ofopinion words according to the psychological Harvard-IV dictionary.Additionally, it allows importing case-specific dictionaries in combinationwith the General Inquirer dictionary to perform affect analysis. It allows togroup data at various levels of time (from day to year) and thus itsupports converting text corpus into time series data. Data range filters canalso be applied on the input. Rocksteady has three output options which includeTotal, percentage and z-score values.

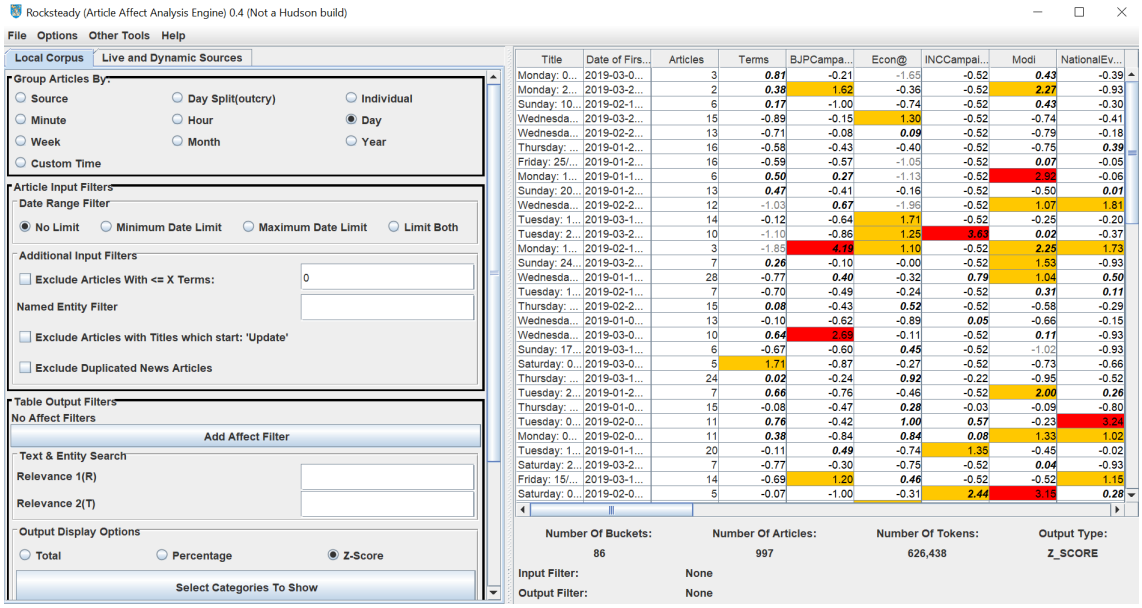

Figure 3.4: Snapshot of Rocksteady User Interface

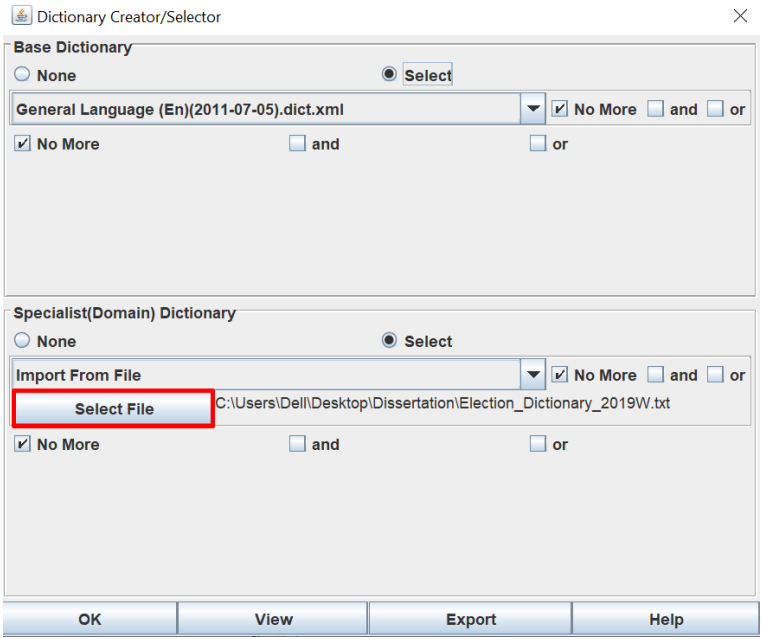

Figure 3.5: A snapshot of UI window to upload domain specific dictionary in Rocksteady

3.5.3 GRETL

Gretl (Gnu Regression, Econometrics and Time-series Library) is an open-source statistical software for econometric analysis. It is used for time series estimation to fit the vector Autoregression model and to estimate the statistical significance of the model results. It takes a time series data involving multiple variables as input and fits a regression model based on the given data using exogenous and endogenous variables. Additionally, gretl also allows to choose the order of the VAR (number of lags) to consider, the variables to include in the model, and whether the model should contain constant, trend, or seasonal dummies. Gretl outputs the result of F-statistic and its p-value, which is used to test multiple hypotheses in a linear regression model, along with its statistical significance. From the main Gretl window, choose Model > Time series > Vector Autoregression to bring up this dialog box. We can set the lag order and add the variables to the box labeled Endogenous and Exogenous Variables along with 'Include a constant' box checked.

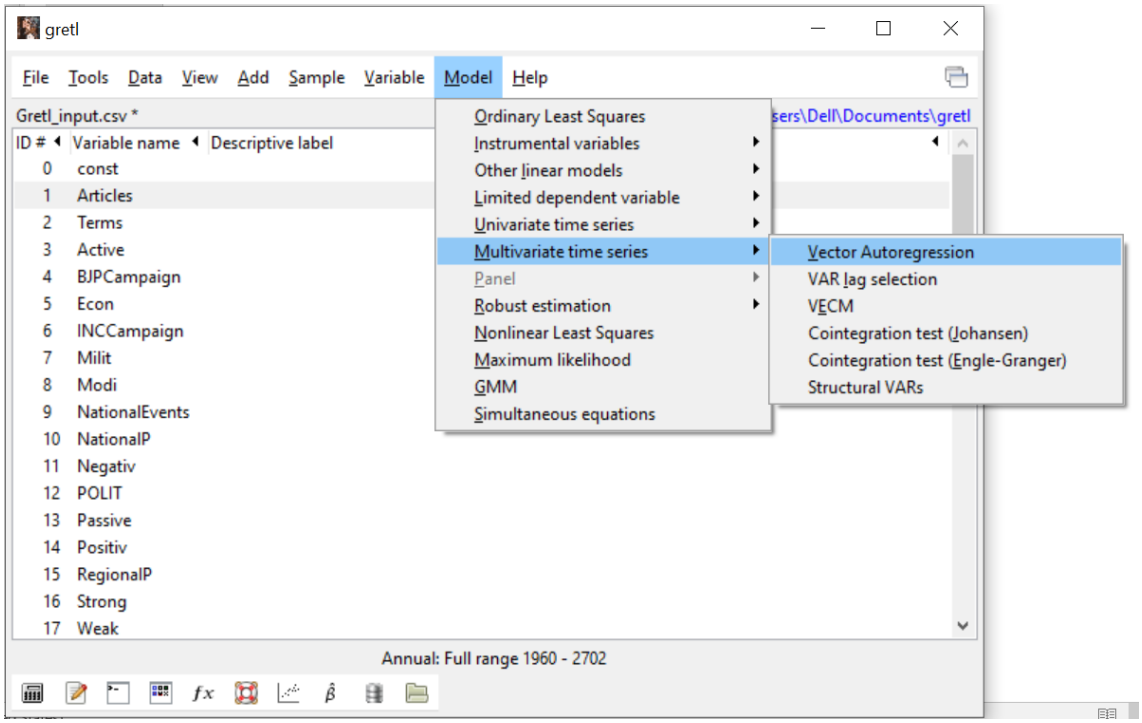

Figure 3.6: A snapshot of Gretl interface illustrating the option flow to create the VAR model to conduct multivariate time-series analysis.

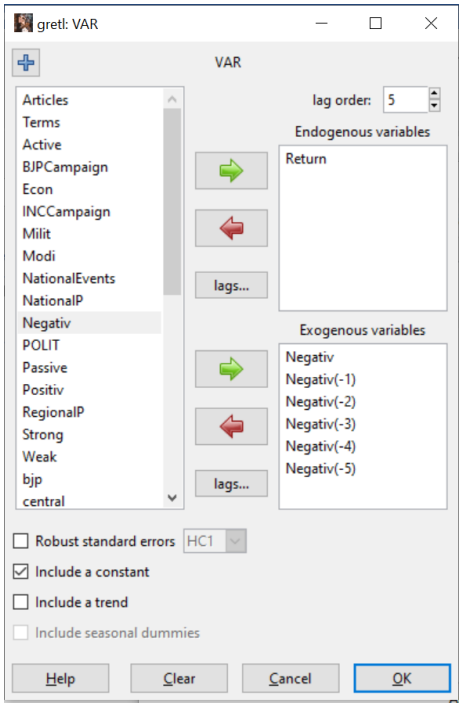

Figure 3.7: A snapshot of the dialogue box to set the endogenous and exogenous variables along with their lag values for VAR modeling in Gretl

3.5.4 Python

Python supports multiple data and statistical analysis libraries for data cleaning, exploration, parsing, statistical summarization and data visualization. Python has been used for various tasks in this case study which include web scraping to collect the keywords for domain-specific dictionary, parsing the collected news media corpus to convert into a rocksteady readable format, data pre-processing, transformation and exploratory analysis of sentiment and financial time series datasets, creation of stylized facts for statistical analysis of time series data and data visualization.

Chapter 4

Case Study - Implementation

India is a democratic country which holds periodic free elections, generally every five years to elect members of its parliament. A Parliamentary system is followed in India, where the leader of the majority party in the parliament is elected as the Prime Minister, holding the highest executive powers in the country. The latest Indian General Election conducted in 2019 is one of the largest electoral exercises in human history which witnessed the participation of approximately 900 million voters. The elections witnessed a highest national voter turnout of 67%. The electoral process for the 17th Indian general elections were conducted in seven phases, from 11 April to 19 May 2019.The results were declared on 23 May, 2019 voting BJP-led National Democratic Alliance (NDA) back to power. The general elections were extended into 2020 for 2 major states and by-elections were conducted in regional states amidst the covid-19 pandemic, consequent lockdowns and economic misery.

4.1 Indian General Elections (2019-2020)

4.1.1 Indian Political Scenario

India is a democratic republic country based on a federal structure of government. India follows a dual polity system which is federal in nature, i.e. there is a central authority at a national level and states hold power at the periphery. The LokSabha or the parliament represents the people of India as a whole, for which periodic general elections are conducted generally every 5 years. India adopts a multi-party system for political parties. As per the latest listing from Election Commission of India, there are about 120 registered parties with 77 national parties, 43 state parties and 621 unrecognized parties ECI (2019).

India is a democratic country with a large number of political parties. The two main parties that dominate Indian politics are the Bharatiya Janata Party(BJP), which is the

leading right-wing nationalist party and Indian National Congress(INC), which is generally considered the center-left in its ideological orientation. India also has two major political party alliances, National Democratic Alliance(NDA) which is a Right-wing coalition led by BJP and United Progressive Alliance (UPA) - Centre-left coalition led by INC. NDA is led by Mr. Narendra Modi and UPA is led by Mr. Rahul Gandhi. INC remains the longest ruling party in the history of Indian Politics by heading the central government for more than 54 years. However, in 2014, INC suffered a massive defeat with less than 20% vote share for the first time in Indian history. At the same time, after a landslide victory in 2014 general elections, BJP emerged as the dominant political party altering the political landscape of India by further reducing the strength of the opposition and other regional parties.

4.1.2 Indian Stock Market

India initiated its open economic policy in 1991 and from then on it had gone through many financial crises, survived major crashes and rebounded to rise as an emerging market. By the beginning of financial year 2018, India was recognized as one of the leading emerging markets with a projected economic growth of 7.7 percent for the fiscal year 2018, until it took a hit by the demonetization and Goods and Services Tax (GST) implementation by the Indian government.

The two major stock exchanges in the Indian stock market are Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE). The BSE is the oldest stock exchange established in 1875, which was recognized by the Indian Government in 1957. As of 2020 about 5,518 firms have been registered with BSE BSE (2021a). For this case study, BSE Sensex is considered as a means to evaluate the stock market performance. S&P BSE SENSEX index (Stock Exchange Sensitive Index) is a benchmark index and a means of measuring the overall performance of the stock market exchange. As per the latest data from the World Federation of Exchanges, BSE stands at the 10th place in the list of major stock exchanges, with a market capitalization of $3.16 trillion when compared to world's largest stock exchange NYSE (New York Stock Exchange) whose valuation is about $26.2 trillion of Exchanges (2019). The Indian stock market also witnessed various crashes in the last few decades, with the latest major crash in February 2020, due to growing global tensions during the covid-19 pandemic.

4.2 Datasets

4.2.1 BSE Stock exchange data

BSE Sensex data was sourced from the Archive Data of the official BSE website. Three years of every day stock exchange data was collected from January 2018 to December 2020, with data missing during Stock market holidays. BSE has about 70 million registered investors as per the latest data published by BSE BSE (2021b). As it can be observed from the below graph, the stock market witnessed a major crash during March 2020, due to growing tensions amidst covid-19 pandemic. It can also be observed that the market witnessed very high fluctuations in the returns around the same time. However, it can also be observed that the market quickly rebounded and there was an increasing trend of stock market close prices by the end of 2020.

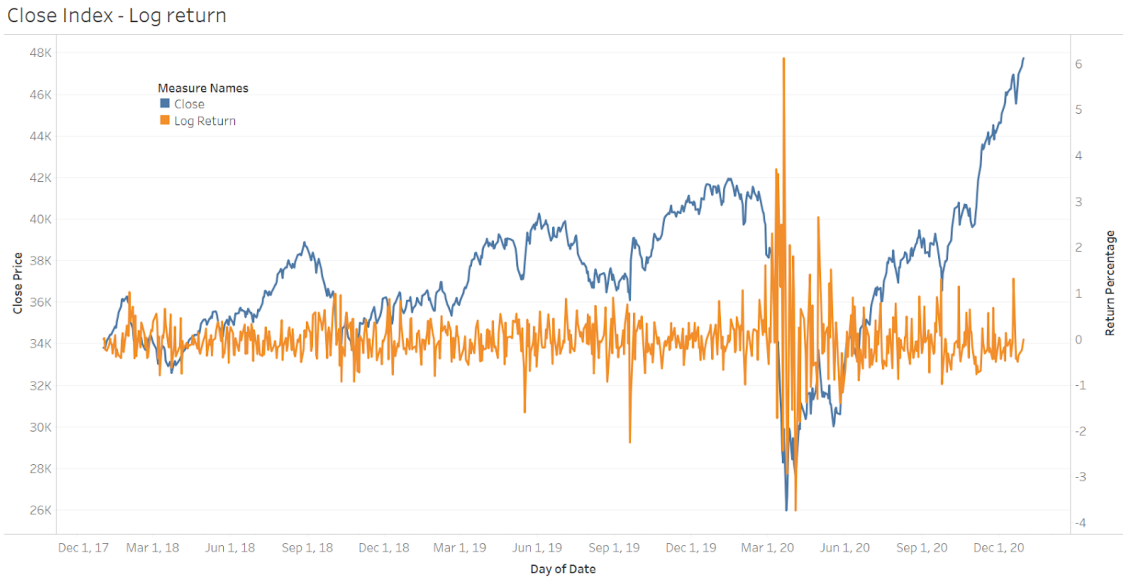

Figure 4.1: Graphical Representation of Log-Returns of BSE Sensex data from Jan 1st,2018 to Dec 31st, 2020

4.2.2 Indian News Articles Corpus

In India, Print media plays a significant role as an empowering tool in the development of the country. It serves as the major source of information and knowledge for the mass population. There is also an embedded socio-cultural routine of receiving a newspaper every morning for a major section of Indian society. History of Indian newspapers dates back to pre-independent times which carried news under British rule NIMC (2021). As of 2018, over 100,000 publications were registered with the Registrar of Newspapers for

India . As of 2018, India has the second-largest newspaper market in the world, with daily newspapers circulation of over 240 million copies Bureau (2021). While the newspaper industry is witnessing a major decline in circulation globally, the print media in India is not only dominant but is also witnessing a major surge in circulation and readership in recent times. By far, Hindi language newspapers have the largest circulation in India. There are various other publications produced in each of the 22 scheduled languages and other spoken languages throughout the country (RNI).While Hindi dailies in India had 186 million readers, English newspaper readership was about 31 million by 2019. Though the English Indian newspapers have a smaller base when compared to Hindi and other regional languages, as per the data released by Indian Readership Survey (IRS), there was an increase of 3 million reader base for English newspapers from 2017 to 2019. As per Audit Bureau of Circulation in 2019, Dainik Bhaskar is the most popular daily newspaper in the country, followed by Dainik Jagran which are published in Hindi (RNI). The Times of India is the third-largest newspaper in India by circulation, largest selling English daily in India.As per latest data (2017-18), the adult literacy rate in India is about 73.2% (RNI). Speaking English is also linked to education, i.e. as per the CMIE survey one third of all graduates could speak English **?**.

Though Hindi is the recognized national language, English still remains the widely used language in India for business, education, bureaucracy, media purposes. As of 2011, 10.67% of the Indian population spoke English. For this case study, Indian newspaper articles are collected from January 2018 to December 2020 to create the text corpus to analyze the sentiment before, during and after the 2019 Indian general elections. 5 out of top 10 Indian English Dailies (**?**) are selected for this case-study. Figure 4.2 presents the graphical representation of the collected newspaper article details. The newspapers analyzed include The Times of India, Hindustan Times, The Telegraph, Indian Express and The Economic Times. The articles are collected using the LexisNexis News Business database provided by the library at Trinity College Dublin and are filtered using the occurrence of the keyword "election". The articles which contain more than 3 occurrences of the word "election" are considered for analysis.

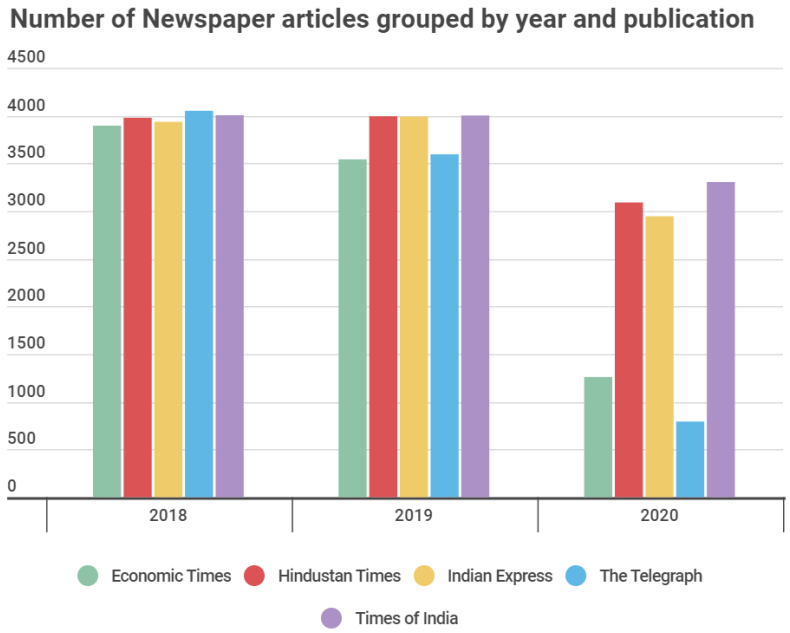

Figure 4.2: A Graphical Representation of the number of articles collected for each newspaper publication for 3 years from 2018 to 2020

4.3 Sentiment in Print Media

With the rise of digital media, the news is now available at almost real-time. Along the same time, there is an increase in the fake news stories questioning the credibility of the news articles. Hence, there is a chance that the news broadcasted on digital and social media contains false and biased information. This can be observed especially in case of political news where the news sites can be highly polarized in favor of a specific political party or candidate and thus can be highly subjective and opinionated. On the other hand, print media had maintained its credibility building trust and reputation over decades. The articles in the print media are generally well curated and audited by professional journalists before being published, and are regarded as reliable sources of information. Thus, print media is chosen as a source to analyze the sentiment.

A number of recent studies show a strong correlation between newspaper/media articles and stock market reactionsBrown and Cliff (2005), Baker and Wurgler (2007). Investors in the stock market generally obtain information from the media to analyze and understand the stock market. Thus, there is a possibility that the sentiment (affect) present in the news articles might change the investor's perception and beliefs, in return affecting the stock market.

National election is a significant event leading to important changes in the political and economic policies of a country. These democratic political events can have an impact

on the stock market due to various changes in the financial regulations and other factors. The relationship between politics and investor behavior depends on the evaluation of political risk, which is defined as the political instability that affects investment values Mosley and Brooks (2014). There is an uncertainty in the period of political change and in particular during national elections Bernhard and Leblang (2002).

This study focuses on analyzing the sentiment in the newspaper articles in the national print media, around the election period to understand if there is a statistically significant correlation between the sentiment in print media and the stock market returns.

4.4 Indian Political Dictionary Creation

A Case-specific Indian political dictionary is created to extract various political features from the text corpus. The features can be broadly classified into 4 major categories which include Named entities of Political Leaders and Parties, Political Campaigns, National events occurred during the election time period and Indian administrative zones.

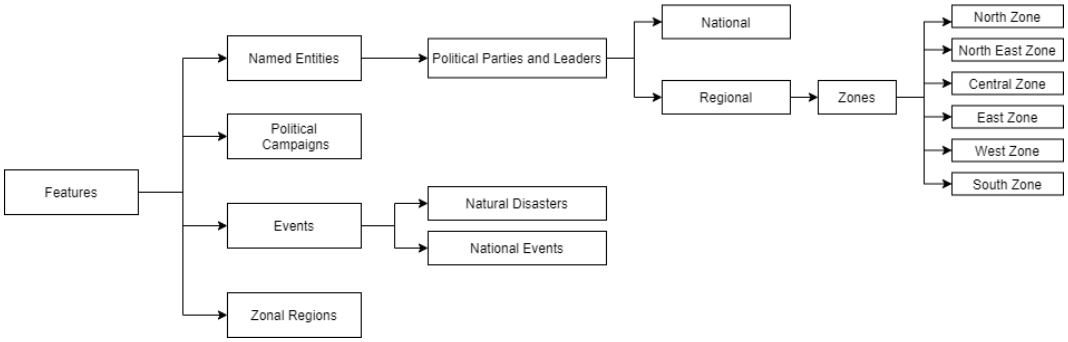

Figure 4.3: Ontology of the created Indian Political dictionary

Figure 4.3 illustrates the ontology of the extracted features and their hierarchy from the collected text corpus. The first feature is the named entities which contain the names of various political parties and their political leaders at the national and regional levels. The regions are further divided into 6 administrative zones. The political leaders and party names are mined from the web and enlisted as entries into the dictionary at the zone level of detail. The major national political parties and their leaders are listed as separate features in the dictionary. The terms from the political campaigns and manifestos of the national parties are mined and enlisted under the political campaign feature category. Various events that occurred before, during and after the election period, including the national events, national movements and the natural disasters such as floods are included under the events feature category. The 6 administrative zones are enlisted as features under the zonal regions category, containing the terms pertaining to each zone. These

feature terms are mined from the collected text corpus as entries into the created case-specific dictionary. These terms are organized in a tab-delimited text file mapping to their corresponding features.

This case-specific dictionary is fed into rocksteady along with the existing general inquirer dictionary to get the calculated affect score. The frequency of feature occurrences in the collected corpus is considered to analyze the sentiment using a bag of words model.

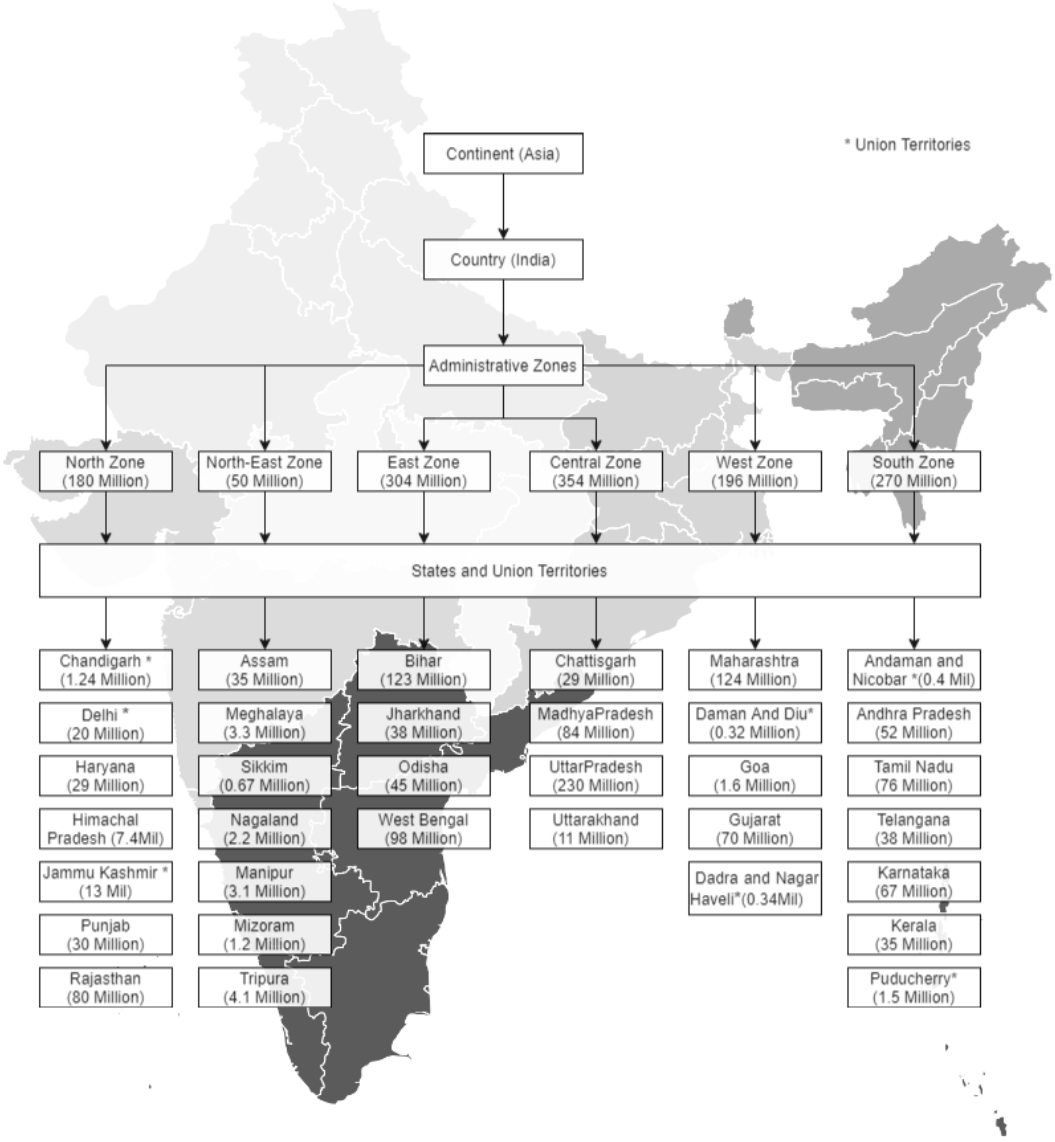

Figure 4.4: Map that outlines the administrative zonal divisions of India as a part of the States Reorganization Act, 1956

The states of India have been grouped into six zones by Indian administration: Central, East, North, North-east, West and South zones. The states, political parties and political

candidates are grouped based on these zone based divisions for the creation of a dictionary.

Feature Category	Features	Feature Description
Affect Terms	Active	Activity (Active)
	Passive	Activity (Passive)
	Negative	Valence Polarity (Negative affect)
	Positive	Valence Polarity (Positive affect)
	Strong	Potency(Strength)
	Weak	Potency(Weakness)
Named Entities	Modi	Current prime minister Mr. Narendra Modi
	Rahul	Opposition party leader Mr. Rahul Gandhi
	BJP	BJP party and its coalition party(NDA)
	INC	INC party and its coalition party(UPA)
Political Campaigns	BJPCampaign	BJP Campaign issues and policies
	INCCampaign	INC Campaign/Election manifesto
Political Parties and Candidates	NationalP	National parties participated in the general election
	centralP	Central zone region
	eastP	East zone region
	northP	North zone region
	northeastP	Northeast zone region
	southP	South zone region
	westP	West zone region
Events	NationalEvents	National Events/Natural disasters occurred during the period
	Covid19	Covid19/Pandemic related terms
Zones	central	Central zone related terms
	east	East zone related terms
	north	North zone related terms
	northeast	Northeast zone related
	south	South zone related terms
	west	West zone related terms

Figure 4.5: The table contains the list of features in the Case-specific dictionary and its respective descriptions in relation to 2019 Indian General Elections

The features analyzed for this case study are broadly divided into six major categories. They are affect terms, Named Entities, Political Campaigns, Political Parties and Candidates, Events and Zones. The affect features include the affect terms such as Active/Passive, Positive/Negative which define the sentiment in the text. The Named entities include the names of leading and opposition party names as well as party leaders to understand the frequency of these terms in the print media.The Political Campaigns include the terms/slogans/agendas of the political campaigns conducted by both the national parties. The Political Parties and Candidates category consists of names of all the political parties and candidates participated in the general elections. The Events category contains the terms related to all the national events including political events, natural calamities and pandemic. The Zones category contains the terms related to each of the six administrative zones in India.

4.5 Summary Statistics

4.5.1 Stock Market Returns

The below plot represents the Sensex log returns as data points and it can be observed the distribution of individual data points within the timeline. It can be seen that fluctuations in the returns are more drastic during April-May of 2020, that is when the first few cases of Covid-19 (Coronavirus) were traced in India and the entire country went into a lockdown.

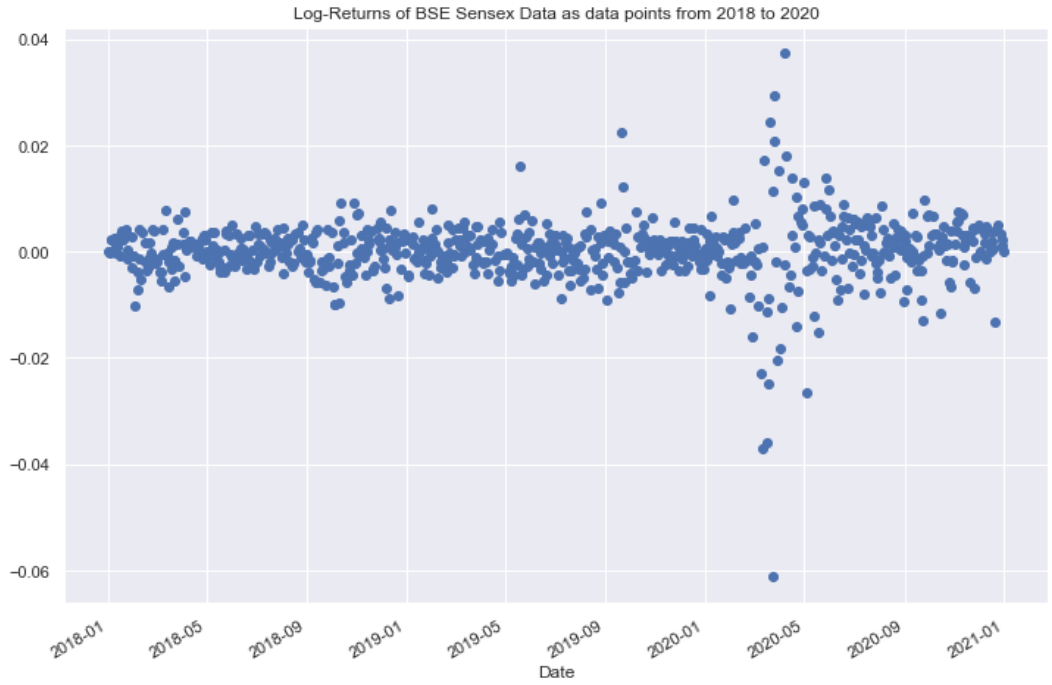

Figure 4.6: Data points representation of BSE sensex Log-returns from Jan 1,2018 to Dec 31,2020

Measure	Sensex Returns
Mean($10^4\mu$)	2.017
Std.Dev(10^2)	0.59
Minimum	-0.06124
Maximum	0.0373
Median	4.44x10-4
Skewness	-1.82
Kurtosis	23.314

From the above table, it can be observed that the values of mean, standard deviation of log returns are close to zero suggesting that there are only minor changes in the return

values. Both the minimum and the maximum values occurred during the Covid-19 pandemic time (Mar-Apr 2020), when the market fluctuations are very high. The minimum return value is -0.0612 which represents the maximum fall in Sensex from the previous day occurred on March 23rd, 2020. The maximum return value of 0.037 represents the maximum increase in Sensex from the previous day which occurred on April 7th, 2020. Skewness is a measure of asymmetry or the deviation from the perfectly normal distribution. Skewness value of -1.82 would suggest that the returns are substantially negatively skewed as the skewness is less than 1. A very large Kurtosis value of 23.314 would suggest that the distribution is a leptokurtic distribution, i.e. the distribution has heavy tails, with large outliers in the data as kurtosis value is >3 (excess kurtosis >0). The above Skewness and Kurtosis values suggest that the distribution of log returns is not a normal distribution.

As it can be observed in the below figure 4.7 the presents the probability distribution of log returns, the distribution is quite centered around zero and it seems symmetric. It can also be observed that the boxplot is full of outliers. The IQR is quite narrow when compared to the total distribution range. This phenomenon is known as fat tails which refers to the probability distributions with relatively high probability of extreme outcomes.

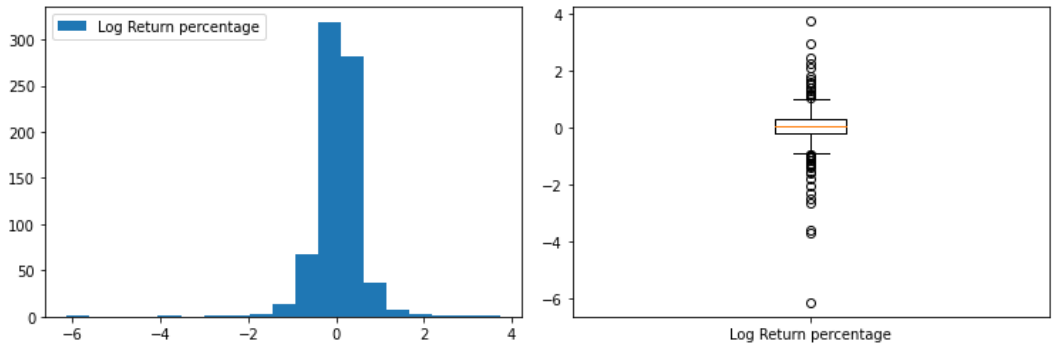

Figure 4.7: Probability ditribution of the log return value

Figure 4.8 illustrates the month-wise seasonal trend in the Sensex returns data. It can be observed from the plot that the median value remains almost constant and close to zero in all the months. It can also be observed that the median value of the log returns almost remains positive in all months except for the months of February and September. The number of outliers in the months of March to May is higher than that of the remaining months in the observed timeline. The log return values in March and October have a wider range when compared to the remaining months.

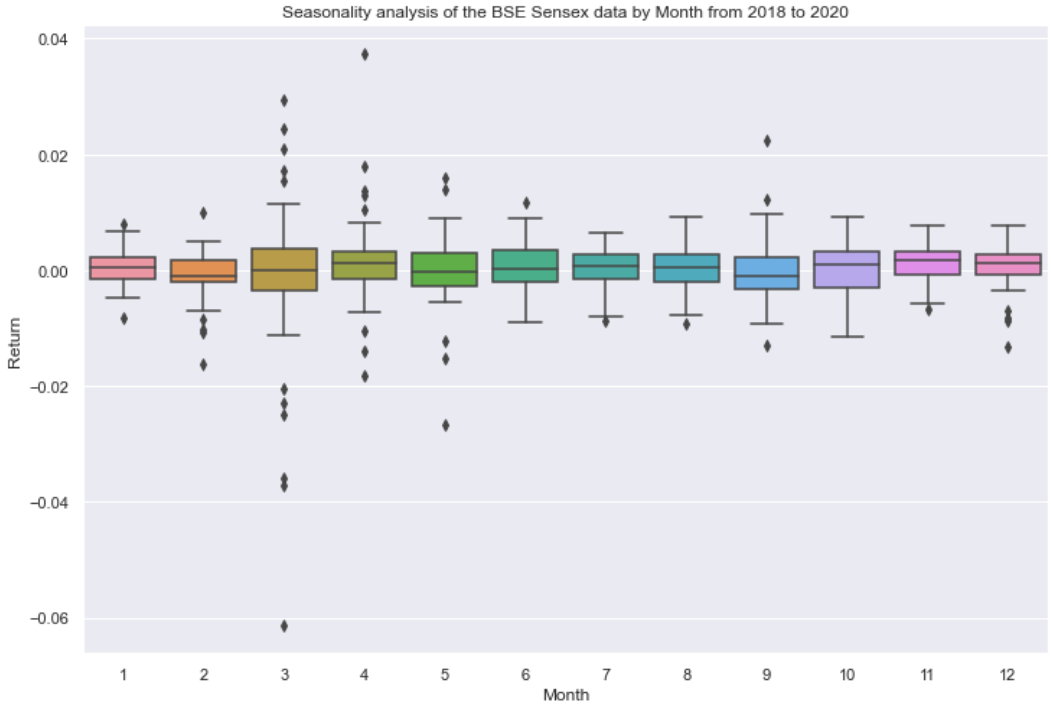

Figure 4.8: Boxplot to analyse the seasonality in BSE Sensex data of 3 years

4.5.2 Sentiment during Indian Elections

Below table shows the summary statistics for various sentiment features (Negative) and other political features extracted using the case-specific dictionary which include BJP-Campaign, INCCampaign, BJP, INC, Modi (Mr. Narendra Modi), Rahul (Mr. Rahul Gandhi), NationalP (National Party), RegionalP (Regional Party), Covid19 (Covid-19 associated terms).

Measure	Sensex Returns
Mean($10^4\mu$)	2.8
Std.Dev	0.9934
Minimum	-2.5519
Maximum	3.9418
Median	-0.0020
Skewness	0.2839
Kurtosis	0.4436

The minimum value of Negative sentiment on media is 1.211 in April 2020 and the maximum negative sentiment went up to 3.56 in May 2020. It can be noticed that there has been major fluctuation of the negative sentiment in the second quarter of 2020, during the covid-19 pandemic, where BSE Sensex had also experienced major fluctuations in its

value. The low values of skewness and kurtosis of negative sentiment indicate that it almost follows a normal distribution.

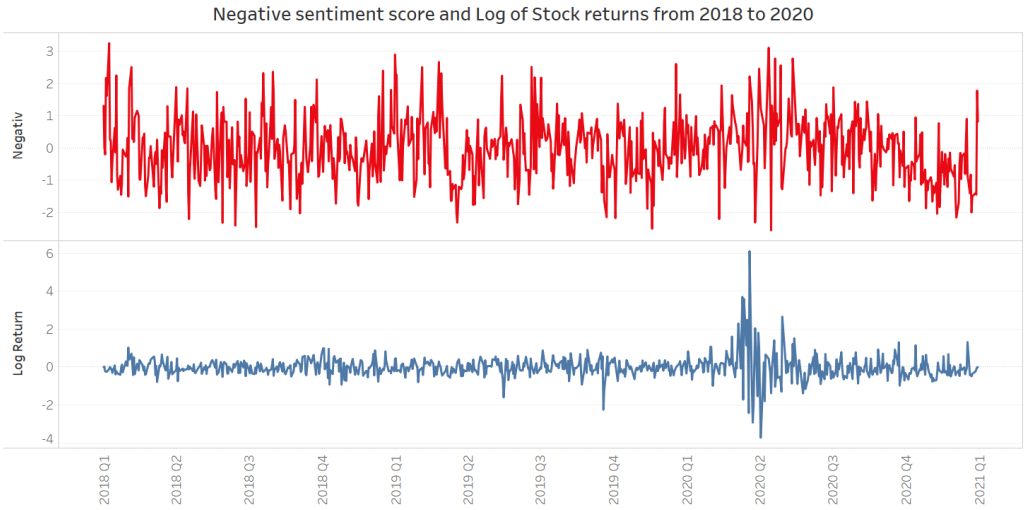

Figure 4.9: An illustration of the Negative sentiment and Log returns from 2018 to 2020

4.5.3 Newspaper Polarity Analysis

This section describes the print media coverage of various newspapers on various political features. Figure 4.10 describes the frequency of occurrence of various feature terms in the newspaper's articles. All the newspapers have consistently higher media coverage of BJP than INC throughout the 3-year duration. The negative scores remain almost similar for the entire duration. A key observation that can be made from the below results is that the frequency of the terms related to the party in power is relatively much higher than that of the opposition. The frequency of the terms related to the governing party (BJP) and the party leader/prime minister (Modi) is significantly higher when compared to that of the opposition party (INC) and opposition party leader (Rahul) consistently among all the newspaper publications.

4.5.4 Feature Correlation Analysis

This section describes the correlation of created political features by conducting a pair-wise study to evaluate the strength of relationship between them. It can be observed from Figure 4.11 that the return value is negatively correlated with the negative sentiment, indicating that the increase in negative sentiment has a negative impact on the stock market. The returns are positively correlated with governing party BJP and BJP-

Newspaper	Year	Articles	Negativ	BJPCamp	INCCamp	Modi	Rahul	BJP	INC	NationalP	RegionalP	Covid19
Economic Times	2018	3884	2.622	0.128	0.006	0.191	0.027	0.711	0.425	0.952	0.204	0.046
	2019	3536	2.500	0.118	0.008	0.198	0.020	0.624	0.331	0.783	0.192	0.057
	2020	1255	2.349	0.144	0.005	0.166	0.016	0.668	0.442	0.946	0.331	0.205
Hindustan Times	2018	3968	2.155	0.107	0.008	0.261	0.056	1.297	0.919	2.102	0.472	0.018
	2019	3990	2.055	0.121	0.007	0.263	0.043	1.273	0.740	1.930	0.614	0.016
	2020	3084	2.026	0.087	0.003	0.141	0.032	0.909	0.565	1.342	0.725	0.398
The Indian Express	2018	3926	2.512	0.145	0.008	0.224	0.039	0.946	0.642	1.427	0.387	0.035
	2019	3987	2.349	0.165	0.010	0.273	0.031	0.991	0.631	1.476	0.413	0.028
	2020	2939	2.459	0.104	0.004	0.136	0.014	0.612	0.456	0.982	0.312	0.335
The Telegraph	2018	4041	2.615	0.187	0.017	0.293	0.030	0.987	0.453	1.159	0.344	0.033
	2019	3589	2.664	0.218	0.012	0.333	0.030	0.980	0.419	1.062	0.273	0.035
	2020	788	2.711	0.202	0.006	0.211	0.015	0.785	0.360	0.903	0.277	0.095
Times of India	2018	3994	2.067	0.078	0.005	0.128	0.035	0.837	0.654	1.410	0.426	0.030
	2019	3998	1.915	0.079	0.006	0.144	0.021	0.812	0.501	1.262	0.507	0.029
	2020	3299	2.010	0.086	0.002	0.088	0.011	0.715	0.458	1.140	0.570	0.344

Figure 4.10: Rocksteady sentiment output grouped by Newspaper, Features and Year

Campaign and the existing prime minister Modi (Mr. Narendra Modi). At the same time, returns are negatively correlated with the opposition party INC and INCCampaign and opposition leader Rahul (Mr. Rahul Gandhi), which gives certain insight into how the political names mentioned in the newspaper articles have a correlation with the stock price returns.

4.6 Statistical Analysis

The main objective of this research is to analyze the impact of negative sentiment and other political features on the stock market price returns, if any. This section presents the statistical models and their results to understand the strength and significance of the above discussed explanatory variables (Sentiment and Political terms) on the response variable (Returns). The statistical modeling is performed in three stages: uni-variate involving a single explanatory variable, bi-variate, which contains two explanatory variables and multivariate, which involves more than two variables to explain the response variable. Various econometric models are created with the returns as the endogenous variable and all other sentiment and political features as exogenous or external variables. The developed models are enlisted in the table in Figure 4.12. A progressive approach is taken by adding a single variable each time to the existing model to analyze the impact of adding the additional variable.

Model 1 is uni-variate where the return value is regressed on itself for five lagged values. Model 2 to Model 7 are bi-variate models which include the lagged values of return along with an additional variable which represents sentiment or political proxies such as negative sentiment (S), BJP (B), INC(I), Regional Parties(RP), National Events(NE) and Covid19(CV). Model 9 to Model 12 extend the previous models by including an additional

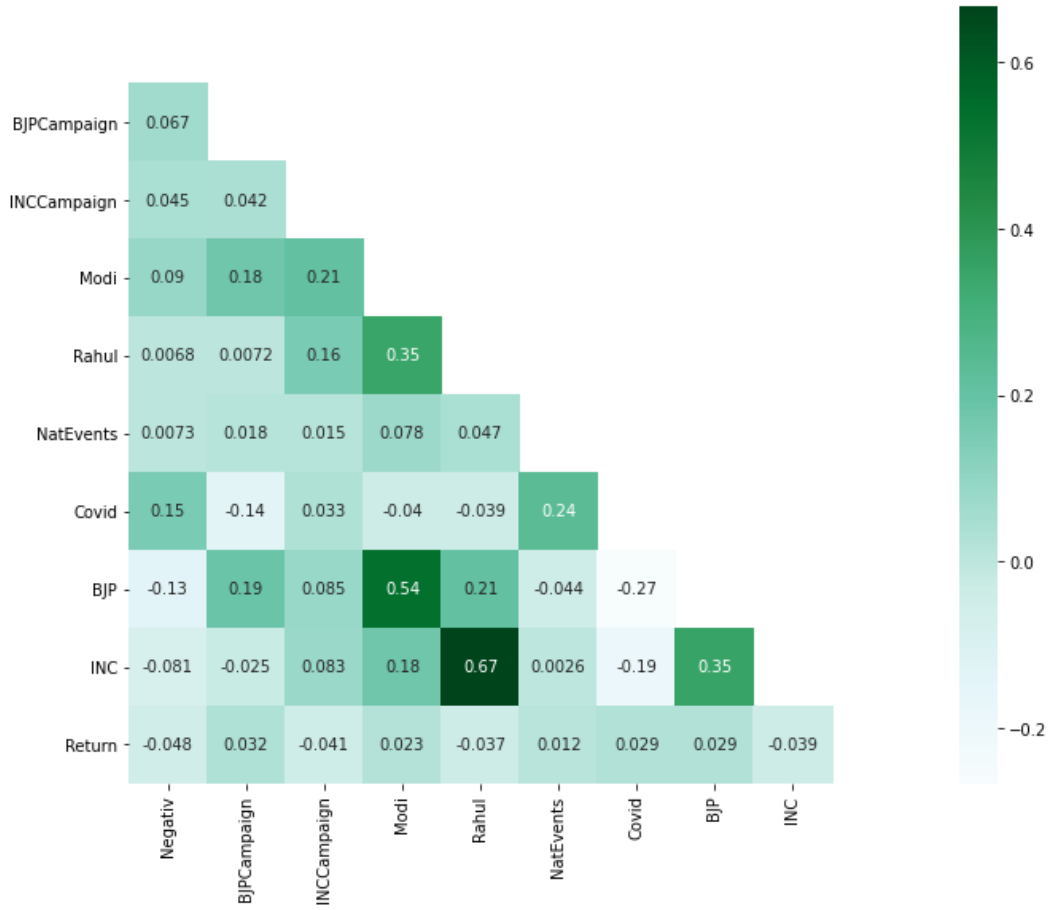

Figure 4.11: Heat map illustrating the correlation between extracted political feature variablesr

	Statistical Model	Exogenous Variable(s)
Model 1	$r_t = C + \alpha L_5 r_t + \varepsilon_t$	-
Model 2	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \varepsilon_t$	Sentiment(S)
Model 3	$r_t = C + \alpha L_5 r_t + \gamma L_5 B_t + \varepsilon_t$	BJP(B)
Model 4	$r_t = C + \alpha L_5 r_t + \theta L_5 I_t + \varepsilon_t$	INC(I)
Model 5	$r_t = C + \alpha L_5 r_t + \eta L_5 RP_t + \varepsilon_t$	Regional Parties(RP)
Model 6	$r_t = C + \alpha L_5 r_t + \mu L_5 NE_t + \varepsilon_t$	National Events(NE)
Model 7	$r_t = C + \alpha L_5 r_t + \rho L_5 CV_t + \varepsilon_t$	Covid19(CV)
Model 8	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \varepsilon_t$	Sentiment(S) and BJP(B)
Model 9	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \varepsilon_t$	Sentiment(S) and BJP(B) and INC(I)
Model 10	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \varepsilon_t$	Sentiment(S) and BJP(B) and INC(I) and Regional Parties(RP)
Model 11	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \mu L_5 NE_t + \varepsilon_t$	Sentiment(S) and BJP(B) and INC(I) and Regional Parties(RP) and National Events(NE)
Model 12	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \mu L_5 NE_t + \rho L_5 CV_t + \varepsilon_t$	Sentiment(S) and BJP(B) and INC(I) and Regional Parties(RP) and National Events(NE) and Covid19(CV)

Figure 4.12: Table containing the list of all the Statistical Models developed using VAR

variable each time to assess the combined effect of the variables on return value.

F-test indicates whether the regression model provides a better fit to the data than a model that contains no independent variables. From the Figure 4.13, it can be observed that the f-test value is statistically significant (p value of f test < 0.05) which indicates that the model's predictions are an improvement over using the average. The results of the F-test performed on the VAR models results as given by Gretl are included along with the model results in the subsequent sections.

	Statistical Model	Null Hypothesis	Alternate Hypothesis
Model 1	$r_t = C + \alpha L_5 r_t + \varepsilon_t$	$\alpha = 0$	*Not* $\alpha = 0$
Model 2	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \varepsilon_t$	$\alpha = \beta = 0$	*Not* $\alpha = \beta = 0$
Model 3	$r_t = C + \alpha L_5 r_t + \gamma L_5 B_t + \varepsilon_t$	$\alpha = \gamma = 0$	*Not* $\alpha = \gamma = 0$
Model 4	$r_t = C + \alpha L_5 r_t + \theta L_5 I_t + \varepsilon_t$	$\alpha = \theta = 0$	*Not* $\alpha = \theta = 0$
Model 5	$r_t = C + \alpha L_5 r_t + \eta L_5 RP_t + \varepsilon_t$	$\alpha = \eta = 0$	*Not* $\alpha = \eta = 0$
Model 6	$r_t = C + \alpha L_5 r_t + \mu L_5 NE_t + \varepsilon_t$	$\alpha = \mu = 0$	*Not* $\alpha = \mu = 0$
Model 7	$r_t = C + \alpha L_5 r_t + \rho L_5 CV_t + \varepsilon_t$	$\alpha = \rho = 0$	*Not* $\alpha = \rho = 0$
Model 8	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \varepsilon_t$	$\alpha = \beta = \gamma = 0$	*Not* $\alpha = \beta = \gamma = 0$
Model 9	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \varepsilon_t$	$\alpha = \beta = \gamma = \theta = 0$	*Not* $\alpha = \beta = \gamma = \theta = 0$
Model 10	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \varepsilon_t$	$\alpha = \beta = \gamma = \theta = \eta = 0$	*Not* $\alpha = \beta = \gamma = \theta = \eta = 0$
Model 11	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \mu L_5 NE_t + \varepsilon_t$	$\alpha = \beta = \gamma = \theta = \eta = \mu = 0$	*Not* $\alpha = \beta = \gamma = \theta = \eta = \mu = 0$
Model 12	$r_t = C + \alpha L_5 r_t + \beta L_5 S_t + \gamma L_5 B_t + \theta L_5 I_t + \eta L_5 RP_t + \mu L_5 NE_t + \rho L_5 CV_t + \varepsilon_t$	$\alpha = \beta = \gamma = \theta = \eta = \mu = \rho = 0$	*Not* $\alpha = \beta = \gamma = \theta = \eta = \mu = \rho = 0$

Figure 4.13: f-test hypothesis for the VAR statistical models

4.6.1 Uni-variate Analysis

Uni-variate Analysis involves regression of a single variable on itself to understand the data and find patterns that exist within it. Table 4.1 shows the result of regression of log return time series data for 5 lags.

$$\text{Model 1: } r_t = C + \alpha L_5 r_t + \epsilon_t$$

where,
r_t indicates the return value at time t
C is the constant
α indicates the coefficient of previous lag return
ϵ indicates the error

From Table 4.1 results, it can be observed that the constant value is very close to zero, which is the mean of the return value. The coefficients show that only the first day and the fifth day lag returns are statistically significant, which means that the correlation between the return price, first day lag return and fifth day lag return is not by chance. The first lag is negatively correlated by a 9% while the later lags nullify the impact by positively correlating with a cumulative of 9%. This indicates that the stock market adjusts itself to maintain an equilibrium. A steady increase or a steady decrease in the returns is an extreme scenario in the stock market.

Constant	0.0002*
r_{t-1}	0.0915*
r_{t-2}	0.0467
r_{t-3}	0.0375
r_{t-4}	0.0288
r_{t-5}	0.1613**
RSS	0.024959
R-squared	0.038790
Adj. R-squared	0.032225
Mean	0.000187
SD	0.005936

Table 4.1: Results of return value regressed on itself for five lags.

4.6.2 Bi-variate Analysis

In the second phase, bi-variate analysis is performed where the stock returns are explained using six other exogenous variables which include negative sentiment (S), BJP (B), INC(I), Regional Parties(RP), National Events(NE) and Covid19(CV) to examine their effects.

$$\text{Model 1: } r_t = C + \alpha L_5 r_t + \beta L_5 r_t + \epsilon_t$$

where,
r_t indicates the return value at time t
C is the constant
α_t indicates the coefficients of return at time t
β_t indicates the coefficient of additional variable at time t
ϵ_t indicates the error at time t
L_5 indicates 5 lags in the time series

From all the bi-variate model results in Figure 4.14 and 4.15, it can be observed that the negative sentiment coefficients from the above results show almost no effect on the return value. The coefficients of negative sentiment (Model-2) are very small in magnitude when compared to the return coefficients and are not statistically significance. Model-3 analyses the BJP party as an exogenous variable along with the log returns. The smaller coefficients of Model-3 suggest that the variable BJP Party only has a small influence on the return value. However, the fourth and fifth day lag values of the variable BJP (B) are observed to be statistically significant. Model-4 includes the variable INC(I) along with the return values and as per the above results, the second and fifth lag values of the INC variable are statistically significant, although the magnitude remains low. Model 5 includes the Regional Parties(RP) as an exogenous variable to explain the return value.

	Model 2	Model 3	Model 4
C	154.48	135.627	156
r_{t-1}	-98987.6 ***	-88825.2 **	-99701.4 ***
r_{t-2}	39039.1	48148	50610
r_{t-3}	33646	36390.5	33448
r_{t-4}	29957.4	25972.3	23755
r_{t-5}	160577 ***	163980 ***	161763 ***
S	-126.305		
S_{t-1}	-345.896		
S_{t-2}	-253.694		
S_{t-3}	29.7		
S_{t-4}	-237.121		
S_{t-5}	61870700		
B		124.901	
B_{t-1}		-18.8847	
B_{t-2}		-27.9117	
B_{t-3}		93.2	
B_{t-4}		573 **	
B_{t-5}		-419.959 *	
I			-211.062
I_{t-1}			45.2
I_{t-2}			-491.246 **
I_{t-3}			-377.081
I_{t-4}			345
I_{t-5}			425 *
RSS	0.024706	0.024673	0.024503
Adj-R^2	0.034114	0.035410	0.042069
F	3.366361	3.459542	3.942397
P value (F)	0.000149	0.000102	0.000014

Figure 4.14: Table showing results of bi-variate statistical models 2-7 where constant(C), returns (r), negative sentiment (S), BJP(B), INC (I), Regional Parties(RP) , National Events(NE) and Covid-19 (CV) ordered by increasing lags.The coefficients are scaled by multiplying 10^6 to increase the readability of the lower values of coefficients. Statistical significance is represented as *** when $p<0.01$; ** when $p<0.05$; * when $p<0.1$

	Model 5	Model 6	Model 7
C	130.246	147.614	149
$\mathbf{r_{t-1}}$	-90794 **	-89496 **	-97351.2 ***
$\mathbf{r_{t-2}}$	46345.6	46790.7	40578
$\mathbf{r_{t-3}}$	37462	32482.8	35669
$\mathbf{r_{t-4}}$	24481.1	30714.9	36810
$\mathbf{r_{t-5}}$	159837 ***	162515 ***	165120 ***
RP	371.847		
$\mathbf{RP_{t-1}}$	-337.327		
$\mathbf{RP_{t-2}}$	-46.7193		
$\mathbf{RP_{t-3}}$	-229.56		
$\mathbf{RP_{t-4}}$	407		
$\mathbf{RP_{t-5}}$	313		
NE		-17.4188	
$\mathbf{NE_{t-1}}$		309.886	
$\mathbf{NE_{t-2}}$		39.6	
$\mathbf{NE_{t-3}}$		-462.713 *	
$\mathbf{NE_{t-4}}$		280.578	
$\mathbf{NE_{t-5}}$		-63.1424	
CV			175
$\mathbf{CV_{t-1}}$			380 *
$\mathbf{CV_{t-2}}$			147
$\mathbf{CV_{t-3}}$			-590.982 ***
$\mathbf{CV_{t-4}}$			-37.0163
$\mathbf{CV_{t-5}}$			-247.157
RSS	0.024644	0.024782	0.024567
Adj-R^2	0.036571	0.031147	0.039580
F	3.543275	3.153939	3.761170
P value (F)	0.000072	0.000348	0.000030

Figure 4.15: Results of Bi-variate models VAR(Continued)

The coefficients in model 5 have very low magnitudes when compared to the return value coefficients and are not statistically significant. Model-6 combines the lag values of returns and the lag values of National Events (NE) to explain the return value and it is observed that the third lag value of NE is statistically significant, which means that the variable has certain but a very weak impact on the return price. The final variable analyzed in the bi-variate analysis is the Covid19 (CV) explanatory variable which tries the analyze the impact of Covid related terms in the news media on stock market returns. It can be noticed that the first and third lag values of the CV variable has an impact on stock return

4.6.3 Multi-variate Analysis

Figure 4.16 and 4.17 shows the regression coefficients for each of the lagged values of the variables used in the VAR model. The statistical significance of the regression coefficient is represented as the probability (p-value) of the null hypothesis that the regression coefficient is used in the equation is equal to 0. Statistically significant values are flagged with stars - One star if a p-value is less than 0.05, two if a p-value is less than 0.01, and three if a p-value is less than 0.001.

	Model 2	Model 8	Model 9	Model 10	Model 11	Model 12
C	154.484	147.192	152.277	148.66	136.11	133.733
$\mathbf{r_{t-1}}$	-98987.6 ***	-95378 ***	-106576 ***	-106293 ***	-107504 ***	-109894 ***
$\mathbf{r_{t-2}}$	39039.1	41735.1	43640.8	43827.3	42208.9	37680.4
$\mathbf{r_{t-3}}$	33646	32845.4	29753.7	31452.9	22232.4	21537.1
$\mathbf{r_{t-4}}$	29957.4	26877.7	21949.3	21903.3	20405.8	26315.9
$\mathbf{r_{t-5}}$	160577 ***	162895 ***	161508 ***	159250***	161175 ***	160459 ***
S	-126.305	-122.042	-123.363	-124.828	-113.059	-107.036
$\mathbf{S_{t-1}}$	-345.896	-349.424	-366.062	-406.394 *	-401.61 *	-433.363 *
$\mathbf{S_{t-2}}$	-253.694	-227.045	-244.806	-254.612	-278.081	-264.143
$\mathbf{S_{t-3}}$	29.7135	68.0086	49.2265	23.3597	25.6996	58.8411
$\mathbf{S_{t-4}}$	-237.121	-211.565	-214.519	-148.661	-145.768	-191.225
$\mathbf{S_{t-5}}$	61.8707	-3.29234	60.4734	125.87	117.998	172.859
B		102.452	214.74	126.729	121.393	96.02
$\mathbf{B_{t-1}}$		-46.6575	-73.9568	63.3002	35.5268	118.183
$\mathbf{B_{t-2}}$		-92.1146	151.799	217.623	243.333	236.732
$\mathbf{B_{t-3}}$		81.8805	182.739	333.038	448.502	396.706
$\mathbf{B_{t-4}}$		543.332 **	498.18 *	432.211	468.982 *	470.458 *
$\mathbf{B_{t-5}}$		-438.013 *	-666.059 ***	-857.926 ***	-861.143 ***	-885.649 ***
I			-292.321	-289.226	-281.772	-265.937
$\mathbf{I_{t-1}}$			73.344	81.8264	55.3495	67.6075
$\mathbf{I_{t-2}}$			-584.263 **	-587.091 **	-604.859 **	-547.976 **
$\mathbf{I_{t-3}}$			-431.027	-451.467 *	-482.605 *	-579.974 **
$\mathbf{I_{t-4}}$			110.77	114.183	145.155	161.297
$\mathbf{I_{t-5}}$			672.181 ***	692.28 ***	704.983 ***	702.428 ***

Figure 4.16: Table showing results of statistical models 2,8 - 12 where constant(C), returns (r), negative sentiment (S), BJP(B), INC (I), Regional Parties(RP) , National Events(NE) and Covid-19 (CV) ordered by increasing lags.The coefficients are scaled by multiplying 10^6 to increase the readability of the lower values of coefficients. Statistical significance is represented as *** when $p<0.01$; ** when $p<0.05$; * when $p<0.1$

	Model 2	Model 8	Model 9	Model 10	Model 11	Model 12
RP				239.445	287.148	322.895
RP_{t-1}				-421.371	-448.196	-387.249
RP_{t-2}				-143.775	-213.576	-204.779
RP_{t-3}				-371.4	-365.772	-434.685
RP_{t-4}				282.83	315.781	343.964
RP_{t-5}				530.265 **	555.984 **	539.135 **
NE					68.4978	20.884
NE_{t-1}					503.487 **	408.396
NE_{t-2}					15.6021	23.8469
NE_{t-3}					-521.355 **	-446.729 *
NE_{t-4}					235.246	292.853
NE_{t-5}					103.592	107.981
CV						186.994
CV_{t-1}						397.909 *
CV_{t-2s}						124.156
CV_{t-3}						-611.78 **
CV_{t-4}						168.699
CV_{t-5}						-252.833
RSS	0.024706	0.024453	0.023843	0.023498	0.023197	0.022825
Adj-R^2	0.034114	0.036073	0.052215	0.058019	0.062138	0.069199
F	3.366361	2.622373	2.765331	2.565311	2.395141	2.336369
P- value	0.000149	0.000379	0.000021	0.000016	0.000016	8.20e-06

Figure 4.17: Results of Multi-variate VAR models(continued)

In this section, Vector Auto Regression - panel approach is built on initial models 1 and 2 which are further extended by adding the newly created political variables at each step incrementally to assess the combined effect of the sentiment and political variables on the stock returns. The positive and negative coefficient values of the lags indicate the self-stabilizing behavior of stock market returns.

The first and fifth day lags of returns are statistical significance values in all the models similar to above bivariate models. One major observation is that the coefficients of the sentiment and other political variables is very less when compared to the coefficients of lag returns. It indicates that the impact of sentiment and other political features is very small and the return values mostly depend on its own lag values.Model 8 demonstrates the existence of statistical significance in Bt4 and Bt-5, indicating that adding the political variable BJP has an impact on the return value. Both the national political party variables are included in Model 9 and it can be observed that the lag value coefficients of both the additional political variables BJP (B) and INC(I) are significant. Interestingly, the coefficient and the significance of lag values of previous variable BJP(B) increased with the addition of the new variable INC(I). The second, third and fifth day lag values of the INC variable remain statistically significant even with the addition of other political variables. It is also noted that the fifth day lag values of both the political party variables BJP(B) and INC(I) remain significant throughout all the models. Model 10 contains political party variables including two national party variables BJP(B) and INC(I) and the Regional Parties (RP) variable and the results show that the fifth day lag of the RP

variable is statistically significant. Model 11 includes the national events variable (NE) to the previous model 10, and it can be observed that the results of first and third day lags of the NE variable are significant. The third lag value of NE continues to be significant even after adding additional variables. Model 12 is the final model in this analysis which includes all the previous values including Covid19 variable (CV). The results show that the coefficients of first and third day lags of CV variables are significant. Interestingly, the variables

On the whole, it is observed that the RSS decreased and the Adjusted R2 value increased with the addition of new variables into the base model. This approach was implemented to verify if the introduction of newly created political variables can have any influence on the stock market returns. But, from the above results it can be clearly observed that the negative sentiment does not have any statistical significant impact in predicting the stock market returns.

However, as per the above results, the created political features exhibit stronger levels of statistical significance over negative sentiment which confirms that there is a potential work that can be carried out in the future by considering political variables to explain stock market returns during elections.

Chapter 5

Conclusions

5.1 Summary

This research is an attempt to understand the impact of sentiment and other political variables expressed in the print media during a political event, i.e. the general elections on the stock market of an emerging economy. The results are obtained by statistically modeling the sentiment and other political features with the stock market returns. The effective market theory suggests that serial correlation in daily stock returns is close to zero. This research investigates the effective market hypothesis to understand the impact of the print media containing political news on investor's mood and thereby the stock markets. During the course of the case-study, a lot of work has been done in understanding the basics of the Sentiment, Politics, Media and stock market domains.

On the technical front, various tools and softwares have been explored to conduct the analysis. The case study involved accessing online databases, formulating queries to collect sample data, data curation, implementing python for data parsing and transformations, MS Excel for data analysis, Understanding the Affect analysis system (Rocksteady) and Statistical modeling using the econometric software Gretl. As a dictionary based sentiment analysis approach is taken for this research, web scraping the election related information from Wikipedia was also experimented to create the custom domain specific dictionary. Various python libraries for data visualization like matplotlib, statsmodels.graphics and seaborn were also explored as a part of this research. The conducted case study produced statistically significant results and a further opportunity to investigate the impact of sentiment in the print media.

5.2 Challenges

One of the main limitations of this study is sourcing only the Indian English newspapers as a proxy to understand the media sentiment. Hindi is the most spoken language in India and the top two Indian newspaper publications, Dainik Jagran and Dainik Bhakar, are published in Hindi. The readership of English newspapers is very less when compared to that of Hindi and other regional newspaper publications. One other limitation of this work is that the newspaper articles are sourced based on a key-term 'election' to extract all election related news articles for analysis. Though it is a good representation of the election related news, the query to extract news articles could be further improved to source more politically rich and meaningful articles to extract sentiment.

The other limitation of this work is the question of whether the news in print media is a meaningful proxy to analyze the stock market. The current work is based on the assumption that affect or sentiment can be measured by calculating the frequency of affect terms in the text based on an affect dictionary. The affect dictionary used in the current research is the GI dictionary developed by [Grieb (1968)]. Whether the terms pre-tagged in this dictionary are still valid and have the same meaning with the change in time needs a thought for the validity of this approach. One other assumption made in this research work is considering news as an explanatory variable for the changes in the stock market. It is assumed that the general public which includes stock market investors who read the news articles are influenced by the sentiment contained in it. The sentiment in the news has an effect on their opinions and decision making, thereby influencing the stock market movements.

One major challenge encountered while implementing the case study was the data acquisition and curation. LexisNexis contains large volumes of news article information and the data from the LexisNexis database had to be downloaded manually in chunks of 500 articles each time due to the maximum download limit of the server. The data curation of close to 50k articles on publication, yearly and quarterly basis was one of the most time-consuming tasks. The other important challenge encountered is getting the results from Rocksteady, as working with such a huge volume of text corpus (close to 50k news articles) requires higher computational resources in terms of both processing and memory. This would often result in system slowdown and crash occasionally. The entire corpus had to be broken into manageable sizes and the output had to be merged again to get the final output.

5.3 Future Work

The current research has a major scope for further improvement and the potential future work is described in this section. As discussed in the above section, the current research focuses only on analyzing the sentiment in English newspapers which is only about nearly 10% of Indian print media circulation and this research can be expanded further to analyze the sentiment in Hindi (Most read language in India) and other regional languages. Currently, there is a similar ongoing research work by Ahmad and Peya Mowar at Trinity College Dublin to understand the Indian stock market dynamics during 2019 elections by extracting sentiment from Hindi newspapers using a "Hindi" language-based affect dictionary. It would be very interesting and insightful to compare and contrast the current results once their work is published.

The domain specific dictionary created for this research can be enhanced further to include lesser known terms associated with each feature category and more related and meaningful features. An elaborate and better dictionary can ensure more efficient analysis of political variables.

Though thirty-five features in total are extracted from Rocksteady (27 political features and 8 affect features), only six out of thirty-five features were used in the final model. Since this is only an experimental study, it can be further enhanced to include other feature variables to assess the strength of those variables and their any statistical significance in influencing the stock market in subsequent research works.

Bibliography

Achar, C., So, J., Agrawal, N., and Duhachek, A. (2016). What we feel and why we buy: The influence of emotions on consumer decision-making. *Current Opinion in Psychology*, 10.

Ahmad, K., Daly, N., and Liston, V. (2011). What is new? news media, general elections, sentiment, and named entities. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 80–88, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ahmed, S., Cho, J., and Jaidka, K. (2017). Leveling the playing field: The use of twitter by politicians during the 2014 indian general election campaign. *Telematics and Informatics*, 34.

Ante, L. (2021). How elon musk's twitter activity moves cryptocurrency markets.

Antweiler, W. and Frank, M. (2006). Do us stock markets typically overreact to corporate news stories? *SSRN Electronic Journal*.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152.

Baker, M., Wurgler, J., and Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2):272–287.

Barber, B. and 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*, 21:785–818.

Bernhard, W. and Leblang, D. (2002). Democratic processes, political risk, and foreign exchange markets. *American Journal of Political Science*, 46:316.

Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *Journal of Computational Science*, 2.

Brown, G. W. and Cliff, M. T. (2005). Investor sentiment and asset valuation. *The Journal of Business*, 78(2):405–440.

BSE (2021a). Key Companies. `https://www.bseindia.com/markets/keystatics/Keystat_Companies.aspx`. [Online; accessed 10-08-2021].

BSE (2021b). Registered Investors. `https://www.bseindia.com/markets/keystatics/KeyStat_ClientStat.aspx?expandable%20=4`. [Online; accessed 10-08-2021].

Bureau, A. (2021). Registrar of newspapers for india. `http://www.auditbureau.org/`. [Online; accessed 11-08-2021].

Cameron, M. P., Barrett, P., and Stewardson, B. (2016). Can social media predict election results? evidence from new zealand. *Journal of Political Marketing*, 15(4):416–432.

Chan, L. K. C. and Lakonishok, J. (1993). Institutional trades and intraday stock price behavior. *Journal of Financial Economics*, 33(2):173–199.

Craig, G. (2004). *The Media, Politics and Public Life*.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Eberl, J.-M., Wagner, M., and Boomgaarden, H. G. (2017). Are perceptions of candidate traits shaped by the media? the effects of three types of media bias. *The International Journal of Press/Politics*, 22(1):111–132.

ECI (2019). List of Political Parties Participated. `https://eci.gov.in/files/file/10989-3-list-of-political-parties-participated`. [Online; accessed 10-08-2021].

Edmans, A., Garcia, D., and Norli, O. (2007). Sport sentiment and stock returns. *Journal of Finance*, 62:1967–1998.

Engelberg, J. E. and Parsons, C. A. (2011). The causal impact of media in financial markets. *Journal of Finance*, 66(1):67–97.

Fang, L. and Peress, J. (2009). Media coverage and the cross-section of stock returns. *Journal of Finance*, 64(5).

Farrell, B. (1987). *The General Election of 1987*. Duke University Press, United States.

Frijda, N. (1988). The laws of emotion. *The American psychologist*, 43 5:349–58.

Garrett, R. (2019). Social media's contribution to political misperceptions in u.s. presidential elections. *PLOS ONE*, 14:e0213500.

Gerber, A. S., Karlan, D., and Bergan, D. (2009). Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52.

Grieb, W. (1968). The general inquirer: A computer approach to content analysis: Philip j. stone, dexter c. dunphy, marshall s. smith, daniel m. ogilvie, with associates. the mit press, cambridge, massachusetts, 1966. 651 pp. plus xx. *Information Storage and Retrieval*, 4:375–376.

Gupta, P. and Panagariya, A. (2014). Growth and election outcomes in a developing country. *Economics Politics*, 26.

Harjule, P., Sharma, A., Chouhan, S., and Joshi, S. (2020). Reliability of news. pages 165–170.

Ho, J. and Hung, C.-H. (2008). Investor sentiment as conditioning information in asset pricing. *Journal of Banking Finance*, 33:892–903.

Josephs, L. (2005). Book reviews: Emotions revealed: Recognizing faces and feelings to improve communication and emotional life, by TC 1" paul ekman. henry holt and company, new york, 2004, 274 pp. *The American Journal of Psychoanalysis*, 65:409–411.

Junqué de Fortuny, E., de Smedt, T., Martens, D., and Daelemans, W. (2012). Media coverage in times of political crisis: a text mining approach. Working papers, University of Antwerp, Faculty of Business and Economics.

Jürgens, P., Jungherr, A., and Schoen, H. (2011). Small worlds with a difference: New gatekeepers and the filtering of political information on twitter.

Kamps, J., Marx, M., Mokken, R., and Rijke, M. (2004). Using wordnet to measure semantic orientations of adjectives.

Keim, D. and Madhavan, A. (1997). Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. *Journal of Financial Economics*, 46(3):265–292.

Kelly, S. and Ahmad, K. (2018). Estimating the impact of domain-specific news sentiment on financial assets. *Knowl. Based Syst.*, 150:116–126.

Keltner, D. and Lerner, J. (2010). *Emotion*, pages 317–352. Wiley, New York.

Levy, T. and Yagil, J. (2011). Air pollution and stock returns in the us. *Journal of Economic Psychology*, 32(3):374–383.

Liu, B. (2012).

Liu, B. and McConnell, J. J. (2013). The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? *Journal of Financial Economics*, 110(1).

Loewenstein, G. and Lerner, J. (2003). *The role of affect in decision making*, pages 619–642. Oxford University Press, Oxford.

Meehan, E. J. (1983). *Reason in Human Affairs. By Herbert A. Simon. (Stanford, Calif.: Stanford University Press, 1983. Pp. viii + 115. 10.00.)*, volume 78.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mitchell, M. and Mulherin, J. H. (1994). The impact of public information on the stock market. *Journal of Finance*, 49(3).

Mittermayer, M.-A. and Knolmayer, G. (2006). Text mining systems for market response to news: A survey.

Mosley, L. and Brooks, S. (2014). Categories, creditworthiness and contagion: How investors shortcuts affect sovereign debt markets. *International Studies Quarterly*, 59.

Mullen, T. and Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. pages 159–162.

Mutz, D. C. (1992). Mass media and the depoliticization of personal experience. *American Journal of Political Science*, 36:483.

NIMC (2021). History of Mass Media. `http://www.nimc-india.com/history-mass-media-india.html`. [Online; accessed 11-08-2021].

Nofsinger, J. (2001). The impact of public information in investors. *Journal of Banking Finance*, 25:1339–1366.

of Exchanges, W. F. (2019). World-exchanges. `https://www.world-exchanges.org/our-work/statistics`. [Online; accessed 10-08-2021].

Petrocik, J. R., Benoit, W. L., and Hansen, G. J. (2003). Issue ownership and presidential campaigning, 1952–2000. *Political Science Quarterly*, 118:599–626.

Phelps, E. A., Lempert, K. M., and Sokol-Hessner, P. (2014). Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience*, 37(1):263–287.

(RNI) (2021a). Registrar of newspapers. `https://rni.nic.in/all_page/pin201718.htm`. [Online; accessed 11-08-2021].

(RNI) (2021b). Registrar of newspapers for india. `https://rni.nic.in/pdf_file/pin2018-19`. [Online; accessed 11-08-2021].

(RNI) (2021c). Registrar of Newspapers (RNI). `https://rni.nic.in/`. [Online; accessed 11-08-2021].

Rogers, E. M., Dearing, J. W., and Bregman, D. (2006). The Anatomy of Agenda-Setting Research. *Journal of Communication*, 43(2):68–84.

Sales, A., Balby, L., and Veloso, A. (2019). Media bias characterization in brazilian presidential elections. pages 231–240.

Santa-Clara, P. and Valkanov, R. (2003). The presidential puzzle: Political cycles and the stock market. *Journal of Finance*, 58:1841–1872.

Sharma, C. K. and Swenden, W. (2020). Economic governance: Does it make or break a dominant party equilibrium? the case of india. *International Political Science Review*, 41(3):451–465.

Shiller, R. (2015). *Irrational Exuberance*. Princeton University Press, 3 edition.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

Symeonidis, L., Daskalakis, G., and Markellos, R. (2010). Does the weather affect stock market volatility? *Finance Research Letters*, 7:214–223.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Tayal, P. and Bharathi, S. V. (2021). Reliability and trust perception of users on social media posts related to the ongoing covid-19 pandemic. *Journal of Human Behavior in the Social Environment*.

Taylor, S. J. (2007). Introduction to asset price dynamics, volatility, and prediction. In *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168.

Tibbitts, C. (1931). Majority votes and the business cycle. *American Journal of Sociology*, 36:596 – 606.

Ullah, R. and Khan, D. (2020). The role of mass media in shaping public opinion.

Ultsch, A. (2008). Is log ratio a good value for measuring return in stock investments? pages 505–511.

Vaishnav, M. and Swanson, R. (2015). Does good economics make for good politics? evidence from indian states. *India Review*, 14(3):279–311.

William, J. B. and Tufte, E. R. (1979). *Public Choice*, 34(1):131–133.

Yang, T., Liu, J., Ying, Q., and Yousaf, T. (2019). Media coverage and sustainable stock returns: Evidence from china. *Sustainability*, 11(8).

Yao, J., Partington, G., and Stevenson, M. (2012). Predicting the directional change in consumer sentiment. *Australian Journal of Management*, 38.

Yu, Y., Duan, W., and Cao, R. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55:919–926.

Abbrevations

BJP - Bharatiya Janata Party
INC - Indian National Congress
UPA - United Progressive Alliance
NDA - National Democratic Alliance
BSE - Bombay Stock Exchange
NSE - National Stock Exchange
NYSE - New york Stock Exchange
S&P - Standard Poor's
NLP - Natural Language Processing
IRS - Indian Readership Survey
GI - General Inquirer
VAR - Vector Auto Regression
RTF - Rich Text Format
IQR - Inter Quartile Range