Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Investigating the Impact of Sentiment on International Financial Markets

Siddharth Gupta
13325731

May 8, 2018

Supervisor: Prof. Khurshid Ahmad

A Dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Declaration

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

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ______________________________ Date: May 6, 2018

Summary

The aim of this study is to investigate the impact of sentiment present in formal media across three major financial markets with different characteristics in the United States, United Kingdom and India. The scope of the work completed can be broken down into three distinct phases: the acquisition and preprocessing of financial data time series and text-based media content; the textual analysis of the corpora of articles collected in order to extract a sentiment time series; and the statistical modelling of daily market return, accounting for a number of independent variables including trade volume and sentiment. Prior to the commencement of work, a review of existing work was conducted in order to determine the best approach to take, and the most appropriate methodologies to employ throughout each phase of the study.

The data acquisition and preprocessing phase sees the collection of historical financial data and articles from formal media sources for each of the three markets, over a time period spanning close to 15 years (September 1st 2003 - January 18th 2018). Over 3,500 observations of daily adjusted closing price and trade volume for indices across the three markets were collected and stored. In order to remove correlations and trends present in the price and volume data, logarithmic return and detrended volume were calculated, where return served as a proxy for financial market performance, and detrended volume was considered as an independent variable that was later used for the modelling of return.

Over 52,000 articles were collected from 25 print and web-based formal media publications in order to compile three corpora for the US, UK and Indian market respectively. Sentiment analysis- a methodology which aims to identify the opinions, evaluations, attitudes and emotions present in written text through the use of natural language processing and text mining techniques- was employed for the analyses of these corpora to extract a sentiment proxy time series. A number of sentiment classification techniques can be used in sentiment analysis, broadly falling under the categories of machine learning approaches and lexicon-based approaches, with this study opting for a dictionary approach under the latter category. Articles in each corpus were represented using a "bag-of-words" model, whereby text was deconstructed into a "bag" of its words, removing context while maintaining word frequencies. Word tokens were then matched against a predetermined glossary of affect-laden words as defined by authoritative sources in order to compute the sentiment present in any given text.

The univariate sentiment proxy time series obtained from sentiment analysis and financial data time series acquired previously was then aggregated into a single multivariate time series. A econometric technique known as vector autoregression was employed for the purpose of analyzing this multivariate time series, modelling return as a function of several independent variables including sentiment and detrended trade volume. Vector autoregression makes use of these independent variables to compute the value of the dependent variable (in this case return), and its linear dependence on these independent variables. This linear dependence calculation, performed using Granger's causality test, yields an estimation of the statistical significance of the contribution toward return.

Visualisation of results was conducted using a panel regression framework for each market, whereby a vector autoregression analysis was first performed on a simple model with few independent variables, and then on subsequently more comprehensive models, with each new model introducing an added independent variable. It is shown that sentiment has a measurable impact on return that is episodic and time-varying in nature, carrying statistical significance across each of the three markets. Furthermore, the models accounting for sentiment exhibit a better fit for the data, and hence are of a higher accuracy, thus justifying the consideration of sentiment as an influencing factor on return.

Given the nature of the time period under investigation- specifically its encompassing of the "Great Recession" from December 2007 until June 2009- a further investigation is conducted to determine how this statistical significance varies across the periods before, during and after the recession. The well-established US market shows that negative sentiment is significant only during the recessionary period. The smaller UK market exhibits significance of negative sentiment during the recession, which is maintained- albeit with reduced contribution- in the period after the recession. The emerging, younger and more regulated Indian market indicates the significance of negative sentiment over all three time periods, although its contribution to return is significantly increased during the recessionary period. Across all three markets, it is shown that considerably better fitting models are obtained when examining the impact of sentiment on return during recessionary periods of high volatility and uncertainty.

Abstract

The 21st century has seen an unprecedented rise in the volume of media content consumed by the general public, disseminated through a plethora of media channels ranging from traditional print-based distribution to modern online distribution on the world wide web. Often generated by influential parties such as formal media sources, the notion of how this information influences the opinions and decisions of those consuming this content has been posited.

This study examines how these formal media sources influence the opinions and emotions of imperfectly rational investors, thus in turn affecting the performance of international financial markets in the United States of America, United Kingdom and India. This is done through the development of a system that leverages natural language processing techniques to extract a proxy for the sentiment present within media articles in the domains of finance, economics and trading. Vector autoregressive methods from econometrics are then employed for multivariate time-series analysis, in order to model the impact of this sentiment on market return.

The system shows that sentiment has a measurable impact on return that is episodic and time-varying in nature, carrying statistical significance across each of the three markets. The US market shows that negative sentiment is significant during recessionary periods, with this significance being washed out after such periods of high uncertainty and volatility. Prior to a recessionary period, the UK market shows no significance of negative sentiment, however this changes upon its onset, and is somewhat maintained through into the period afterward. Finally, the Indian market consistently indicates the significance of negative sentiment, although its contribution is shown to heighten considerably during a recessionary period, thus suggesting a higher and continuous degree of uncertainty present in the more recently established emerging market.

It is observed that statistical models that account for sentiment exhibit a better fit for the data than the models that do not- thus suggesting that the consideration of sentiment as an influencing factor on return is justified and adds value to the modelling of financial return. Furthermore, this fit further improves when applied during a period of recession, indicating that perhaps sentiment is predominantly influential during periods of high market uncertainty and volatility.

Acknowledgements

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

First and foremost, to Professor Khurshid Ahmad, a guiding light who has driven me to challenge myself and get the most value from this work as possible. Thank you for always taking the time to offer your expertise, advice and guidance over the course of this project. Your passion for the field is intoxicating, and has no doubt sparked my curiosity and instilled in me a deep appreciation of the subject. A special thanks should also be made to Dr. Mike Brady, the Computer Engineering M.A.I. coordinator in the School of Computer Science and Statistics.

Thank you to my friends and classmates whom I have interacted with closely over the course of the year. I have taken away a lot from our discussions, and hope the same has applied to you. In particular I would like to thank my good friend and classmate Darragh McKay, who has been a source of valuable advice and support both over the course of the project and my time in university.

Finally, a big thank you to my family- my mum, dad and brother, who have supported me with love, care and encouragement. Thank you for never sparing an expense when it comes to my education and personal development- no doubt this will stand to me for the rest of my life. You are the reason I strive to fulfill my potential, and everyday you inspire me to be the best person I can be.

Contents

List of Figures

List of Tables

Nomenclature

API	Application Programming Interface
BP	Basis Point
BSE	Bombay Stock Exchange
DJIA	Dow Jones Industrial Average
EMH	Efficient Market Hypothesis
ESM	Emerging Stock Market
FTSE 100	Financial Times Stock Exchange 100 Index
GDP	Gross Domestic Product
GRETL	GNU Regression, Econometrics and Time-Series Library
GUI	Graphical User Interface
IIFL	India Infoline Limited
NASDAQ	Nasdaq Stock Market
NBER	United States Bureau of Economic Research
NLP	Natural Language Processing
NYSE	New York Stock Exchange
RSS	Residual Sum Squared
SENSEX	Standard and Poor's Bombay Stock Exchange Sensitive Index
SO-CAL	Sentiment Orientation Calculator
S&P 500	Standard and Poor's 500 Index
SSEC	Shanghai Stock Exchange Composite Index
SVM	Support Vector Machine
UK	United Kingdom
US	United States of America
USA	United States of America
VAR	Vector Autoregression

1 Introduction

The late twentieth century saw technological breakthroughs in the finance industry, with the commencement of the electronic stock exchange- and with it, the inception of electronic trading. From a physical location where traders met and negotiated, the stock market had now expanded into the digital dimension- eliminating previously existing barriers to trading associated with brokers and geographical location, by facilitating the buying and selling of shares anywhere at any time through online access to the stock market [1]. Stock markets thus became significantly easier to access by investors, leading to a considerable increase in the volume of transactions made on the market, and an explosive growth in the availability of important data such as stock prices and volumes.

The creation and deployment of the internet largely facilitated these breakthroughs, but naturally had a diverse range of applications beyond the boundaries of the finance industry. News agencies brought content online, allowing for larger volumes of content to published. Whole new platforms generating vast amount of information such as social networks and blogs were created. Democratization of the technology facilitated widespread public access to this information, thus resulting in an ocean of media content from both mainstream media sources and individual users alike. It enabled members of the common public to voice their opinions on a global stage exposed to the rest of the world.

A reasonable question to ask is how this plethora of uncensored, easy-to-access information influences the opinions (sentiment) and decisions of those consuming this content. Specifically within the finance industry, it is crucial to understand how this information affects market/investor sentiment- the overall attitude of investors toward a particular security or financial market- and in turn, how market sentiment impacts the perceived value of financial assets.

Many studies have performed investigations into the underlying contributors to the current value of financial assets (see Chapter 2 for examples). A brief examination of these studies suggests that the commonly-held belief that financial return is influenced by its historical trend is in fact true, but naive when considered in isolation. Green and Pearson assert that in a dynamic environment such as the stock market, the movement of price or return is determined by the combination of ongoing trends, patterns hidden in charts or graphs, the sentiment in the market, and other unforeseeable events of the future, which could either be political or fundamental issues such as interest rates, RPI, GDP, etc. [2].

This study specifically aims to focus on the area of market sentiment expressed through written media sources, on a daily basis in emerging markets. Many experts argue that all the information regarding any economic entity is reflected in the monetary value of that entity. However, a counter-argument has been posed- that in periods of uncertainty and high-volatility, connotative use of language may be linked to an attempt by the stakeholders in the market to discover the actual value of a share- which suggests that the sentiment of market stakeholders is influenced by what they read or hear in media [3]. This study supports this counter-argument, specifically focusing on how textual media contributes to market sentiment to influence the value of financial assets. Like prior studies, this study employs a technique known as sentiment analysis to identify and extract market sentiment from a variety of media sources.

1.1 A Note on Sentiment Analysis

Also known as opinion mining or emotion artificial intelligence, textual sentiment analysis aims to identify and extract the subjective content of written work through the analysis of the author's opinions, evaluations, sentiments, attitudes and emotions. Techniques such as text mining and natural language processing (NLP) are employed in the analysis in order to aid in the extraction process [1]. There exists a multitude of sentiment classification techniques, but broadly these can be categorized as being of a machine learning approach or lexicon-based approach to sentiment analysis. Figure 1.1 provides a hierarchical overview of these techniques under the two approaches [1].

This study will employ the dictionary-based approach under the latter category, commonly used in previous studies such as those performed by Ahmad et al. and Kelly [3] [4] [5] [6]. This approach makes use of a predefined dictionary of words associated with a specific sentiment polarity strength. Emotions expressed by the author of a text- such as happiness, sadness or depression- can be determined by comparing words in the text against terms from the employed dictionary [1]. Texts present in the corpus to be analyzed are represented using a "bag-of-words" model [7], in which they are broken down into single word tokens, thus removing context, but taking care to preserve term frequency. Though the term "corpus" is used throughout this study, it should not be confused with a corpus approach to sentiment analysis, which attempts to find co-occurrence patterns of words to determine their sentiment whilst maintaining context. Here the term corpus is simply used to reference the collection of media texts gathered in order to perform dictionary-based sentiment analysis.

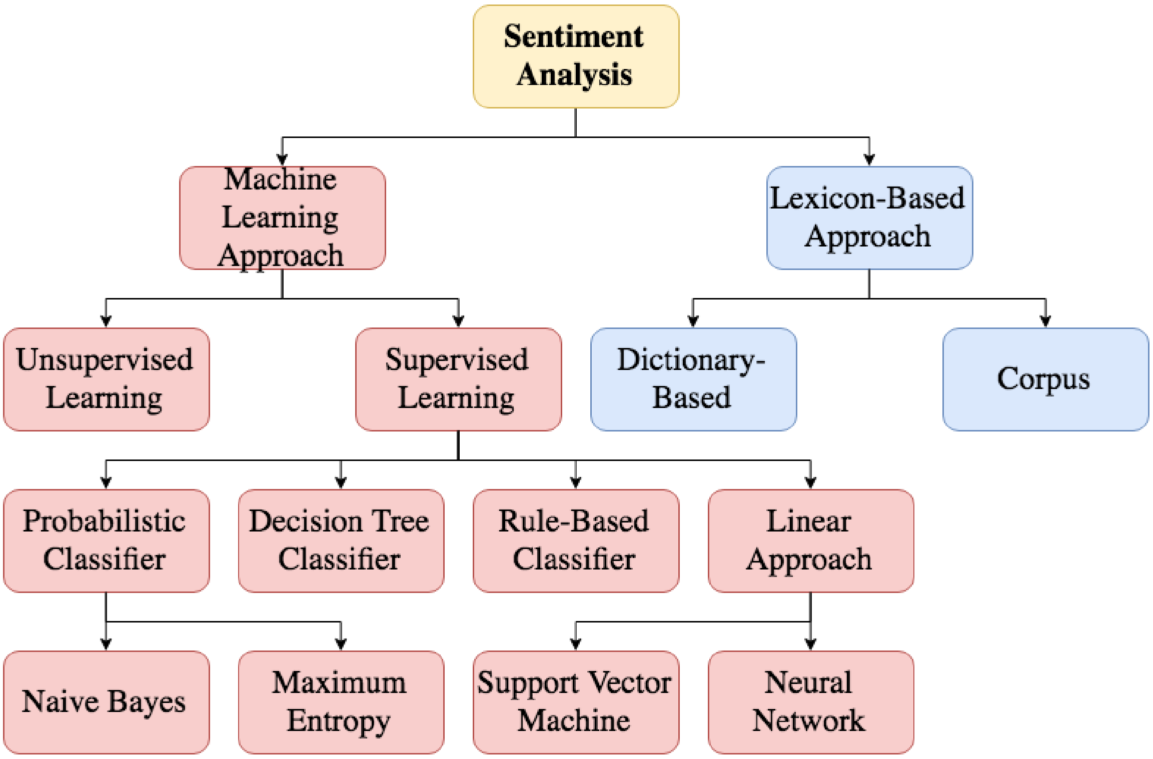

Figure 1.1: Sentiment Analysis Classification Techniques

A significant challenge associated with a sentiment analysis approach to stock market analysis is the notion of duplicity. By duplicity, this study refers to opinion-expressing words that can be perceived as either positive or negative depending on the context it is used in. For example if the word "short" is used for the latency time of some system, then it is considered to have a positive connotation. On the other hand, if it is used to describe the battery life of a mobile device, it then takes on a negative connotation. This challenge is particularly applicable when a domain-specific study such as this is being conducted, as it is common for words to take on different meanings across different domains. Throughout the study, care is taken to ensure that appropriate mitigation measures are taken to address this challenge. Detailed descriptions of these measures are provided over the course of this document upon their introduction.

1.2 Objectives

The aim of this study is to investigate the impact of sentiment present in formal text-based media on both well-established and emerging stock markets, which is accomplished through the pursuit of four main objectives. The first is to analyze a corpus of texts gathered from reputable media sources specific to each market, in order to identify the presence of sentiment and quantify it in the form of a proxy variable time series. The next objective is to build statistical models for financial market return- taking into consideration a number of different influencing variables- that can be used to estimate the statistical significance and contribution of sentiment towards market return. The third is to apply these models to both the established and emerging markets under consideration, facilitating a comparison and contrast of the results across the different markets. The final objective is to apply the models to each market under varying market conditions, in order to investigate how the results differ across periods of economic stability versus periods of economic volatility and uncertainty.

1.3 Motivation

The motivation behind this study is to investigate how techniques in the areas of sentiment analysis and multivariate time series analysis can be jointly applied to real-world models of financial return. Three financial markets have been selected for study: those in the United States, United Kingdom and India. The US market has been selected due to its long establishment, stability and liquidity. The UK market exhibits similar characteristics, albeit it smaller than the US, and additionally serves as an important financial centre in the European Union. The Indian market has been selected for its contrasting characteristics as a relatively recently established market that is part of an emerging and heavily regulated economy. The indices selected for study to serve as proxies for the three markets are selected based on their frequent usage as proxies in previous studies (as discussed in Chapter 3). The formal media publications selected as sources for article data for each market are shortlisted under the criteria of publications with the highest circulation and readership for that market.

As discussed previously, a number of different classification techniques exist for the purposes of identifying and extracting the sentiment present in the collected articles. A dictionary-based approach to sentiment analysis together with the "bag-of-words" model of text representation is employed for this study due to the extensive work that has already been performed in the development and refinement of lexicons that have been made available for public use. The univariate sentiment proxy time series obtained from sentiment analysis and financial data time series acquired previously will then be aggregated into a single multivariate time series. Hence, for the purpose of analyzing this multivariate time series, an econometric technique known as vector autoregression is employed, which will facilitate the modelling of return as a function of sentiment- thus revealing any dependency that may exist between return and sentiment.

1.4 Key Findings

The results obtained from the system indicate that the proxy for negative sentiment obtained from sentiment analysis always carries statistical significance in some form, however has a varied contribution across the three markets. The US market exhibits its significance only on the third day lag term, negatively contributing 1.35 basis points towards return. The UK market on the other hand indicates significance of the proxy on the second and fifth day lags, positively contributing 2.31 and 3.07 basis points respectively. Finally, the Indian market exhibits its significance on the first and third day lag terms, positively contributing 2.20 and 2.64 basis points respectively. The UK and Indian markets additionally suggest that the "same day" term carries significance that- unlike the positive contributions of its lagged counterparts in both markets- negatively contributes a significantly larger amount towards returns (11.78 and 6.65 basis points respectively). Importantly, the models accounting for sentiment exhibit a better fit for the data, and hence are of a higher accuracy, thus justifying the consideration of sentiment as an influencing factor on return.

Further investigating the episodic, time-varying nature of sentiment, results were obtained for the sub-periods before, during and after the recession that took place from December 2007 until June 2009 [8]. In the period prior to the recession, the proxy variable for negative sentiment carries significance only in the Indian market, on the "same day" term, negatively contributing 9.62 basis points towards return. During the recession, negative sentiment comes into significance in the US market on the third day lag term, negatively contributing 5.61 basis points to return. Similarly, the UK market sees the onset of significance for the proxy on the "same day" term, negatively contributing 31.85 basis points. Finally, the Indian market is shown to maintain significance for negative sentiment on the "same day" term, however its negative contribution has increased to 24.65 basis points. Additionally, the first day lag term comes into significance, positively contributing 10.52 basis points towards return. It is further noted that the "goodness of fit" of the model for each market is shown to increase significantly when compared to the period prior to the recession.

Following the recessionary period, the US market loses the significance it had carried on negative sentiment during the recessionary period. Significance is still maintained in the UK market, and although it's contribution is lessened, it extends from what was previously just of the "same day" term (whose negative contribution decreased to 19.09 basis points) to also include the first, third and fifth day lags, positively contributing 4.37, 3.45 and 6.01 basis points respectively. Negative sentiment is shown to maintain its significance in the Indian market, with a negative contribution of 2.11 and 1.65 basis points on the "same day" and fourth day lags respectively, and a positive contribution of 1.54 basis points towards return on the third day lag term. It is noted than an examination of the R^2 terms for the model across the three markets shows an overall reduction in "goodness of fit" compared to the period during the recession. A quick examination of the R^2 terms over all three sub-periods indicates significantly better fitting models during the recession, thus suggesting that sentiment is predominantly influential during periods of high market uncertainty and volatility.

1.5 Thesis Structure

The scope of the work to be completed by this study can be divided into three distinct phases: the acquisition and preprocessing of financial time series data and media content, textual sentiment analysis of the constructed corpora to extract proxy time series for sentiment, and finally the statistical modelling of the financial and sentiment time series. For this reason, several chapters in this study are divided into three sections, with each section focusing on one of these three project phases.

The following chapter offers a review of the methods employed by previous studies in the fields of sentiment analysis and its application to the analysis of financial markets (Chapter 2). Next, an outline of the steps taken by this study throughout each phase of the project is provided, including justification for the selection of particular methodologies in each stage (Chapter 3). The results obtained from each phase of the process are then presented, taking care to ensure that the data provided offers an accurate and complete representation of the outcomes of the study (Chapter 4). Also offered is an analysis of these results, and a discussion of their implications in the context of this study and previous studies in the field. To conclude, the scope of the work performed and main findings of the study are summarized, and suggestions for future work in the area are provided (Chapter 5).

2 Background and Related Work

The efficient market hypothesis (EMH) states that the prices of financial assets fully reflect all available information, and that on the average, competition will cause the full effects of new information on intrinsic values to be reflected "instantaneously" in actual prices [9] [10]. First posited by Fama in 1965, the EMH suggests that assets are always perfectly priced, making it impossible to "beat the market" by purchasing undervalued stocks or selling stocks at inflated prices. The theory thus implicitly argues that noise traders[1] are irrelevant in asset pricing, because trading activity is always performed by the fully-rational, risk-averse arbitrageurs that drive prices close to the fundamental values of assets at the equilibrium position [11].

In a pioneering study, De Long et al. contrasts this idea of perfect rationality, instead suggesting that investor sentiment plays a significant role in the asset pricing process [14]- a view that is shared by several other studies including those discussed later in this section. The study particularly emphasizes the role of noise traders in discovering asset price equilibrium [11]. This suggestion that investors are not perfectly rational decision makers, and in fact take on social, cognitive and emotional biases [15], is the foundation of the relatively new field known as behavioural finance. Behavioural finance seeks to combine behavioral and cognitive psychology with conventional economics and finance to provide explanations for why people make irrational financial decisions [16]. Sentiment analysis is a tool that is used in the field in order to better understand the aforementioned emotional biases, through the extraction and quantification of emotions and opinions present in media.

This study employs sentiment analysis in a textual capacity, in order to extract opinions present in formal media expressed towards financial markets. This sentiment is then jointly examined with financial data in an effort to investigate any potential influence it may have on the performance of the markets in question. As outlined in the introduction, there are three phases of this study: data acquisition and preprocessing, sentiment analysis and

[1] Here the term noise traders refers to irrational investors whose demand is influenced by beliefs or sentiments that are not fully justified by fundamental news [12], perhaps due to a lack of access to *all* information contributing to the intrinsic value of a financial asset [13].

statistical modelling. For this reason, the following review of previous work conducted in the field has been divided into these three sections, with each section providing an overview of the techniques used by previous studies in that specific area.

2.1 Data Acquisition

This section discusses the retrieval of financial and media data, first identifying the information content gathered by previous studies in this area, and then providing an overview of the most common acquisition techniques.

2.1.1 Content

Previous studies examining the influence of sentiment present in media on financial markets are extensive, and span across a plethora of markets across the globe. Analyses performed on each of these different markets necessitates the use of different financial data and media content sources. Perhaps the most examined market is the US financial market- the largest in the world today. Fisher and Statman, Zhao and Ahmad, and Tetlock all examine this market, yet take slightly different approaches in their data retrieval process. With regard to financial data, the former two make use of the S&P 500 index as a proxy for the US market, whereas Tetlock employs the Dow Jones Industrial Average as a proxy. Tetlock relies on the Wall Street Journal as its sole source for its media corpus [17]. Zhao and Ahmad opt to instead build their corpus from a variety of media sources [18] [19]. Finally, Fisher and Statman circumvent the need to build a corpus and perform sentiment analysis, by directly accessing sentiment data for the US market made available by Merrill Lynch [20].

The smaller UK market, exhibiting similar characteristics of development, liquidity and regulation is also examined, albeit with considerable less frequency. Studies performed by Nikkinen and Vähämaa and Gillam et al. both study factors that influence the market, employing the FTSE 100 index as a market proxy [21] [22]. Like Tetlock, Gillam et al. relies on a single source of media content, building a corpus of articles published on the Reuters news agency website.

The vast majority of existing studies focus on developed and mature markets such as the US and UK- countries with stable democratic governance and high GDPs per capita. Fewer studies explore emerging markets such as those present in India and China- less mature, developing countries exhibiting signs of rapid growth and high volatility. Most of those to have done so have been conducted only within the last five years. Zhao and Ahmad extend their studies to the Chinese market, contrasting the effect of sentiment on that market with the effect on the US market. Their exploration of the Chinese market sees the use of the

Shanghai Stock Exchange Composite (SSEC) Index as a proxy for the Chinese market [19]. Once again, they build their corpus from a variety of media sources, however it is important to note that in each case, media sources appropriate to the region being examined are selected. Bhardwaj et al. examines the effect of sentiment on the Indian market for stock market prediction purposes [1], while Kumari and Mahakud investigate whether or not sentiment can predict asset volatility in the Indian market [11]. Interestingly, both studies employ historical prices of the SENSEX and Nifty indices as sources of financial data for the Indian market, thus suggesting these indices serve as effective proxies for the Indian market. Dash and Maitra on the other hand use the Nifty index together with the NSE Midcap and Smallcap indices, in an effort to ensure small, medium and large companies are all encompassed within their study, which aims to determine if sentiment affects stock returns [23] [24].

2.1.2 Retrieval Method

A number of different data retrieval techniques have been employed by previous studies, however they all have two common goals: to obtain the required financial data for the time period being examined, and to obtain the relevant textual media data for that same period.

Gathering the relevant financial data is a trivial task due to its wide availability via third-party services and plugin software packages. For example, studies conducted by both Kelly and Hochreiter make use of the popular Quandl service as a source for financial data [5] [25]- a package that offers both an API and libraries for common analysis tools such as R, Python and Excel.

The retrieval of text-based media information can be broadly broken down into three categories: web scraping of the websites containing the required data, direct access to an application programming interface (API) offered by the source of the data, and utilizing third-party services that gather all the required data automatically and make it available to the user. Studies such as those performed by Kelly and Bhardwaj et al. approach the media content gathering process through web scraping. Both studies make use of Python with the "Beautiful Soup" and "Scrapy" libraries designed specifically for web scraping [1] [5]. These libraries facilitate the parsing of website information and the extraction of this information as a nested data structure. The data obtained must often be put through an extra preprocessing step to organize into a format suitable for analysis. Hence web scraping is more frequently used in cases where the required data is not easily accessible via API or third-party service. A more straightforward approach to the retrieval process can also be employed, directly accessing the required data from the provider of the information via APIs where available. This data comes formatted with a coherent structure throughout the data, thus requiring less pre-processing and simplifying the retrieval process. In both cases however, extending the acquisition beyond a single source often necessitates added work, as different data sources often require different scraping techniques or API access mechanisms.

Studies conducted by Kelly, Zhao and Ahmad perform sentiment analysis on multiple different markets for the purpose of either aggregating or contrasting the obtained results, and thus consider several publications across these markets. These studies build text corpora using Lexis Nexis- a third-party aggregator service providing large amounts of data spanning across multiple published analogue and digital media sources around the world [5] [18] [19]. The use of an aggregator service such as Lexis Nexis circumvents the aforementioned problem of requiring different approaches to obtain data across multiple sources via web scraping or APIs.

2.2 Text Analysis

Within the scope of this study, text analysis refers to the examination of written media, with the aim of identifying and quantifying the sentiment expressed by the authors. As discussed in Section 1.1, sentiment classification techniques can be broadly categorized as falling under either a machine learning approach or lexicon-based approach (see Figure 1.1).

2.2.1 Machine Learning Approach

Advocates of the machine learning approach typically make use of a corpus of annotated text that is used in conjunction with an algorithm(s) that can learn the "affect" of a specific term or text document in general [5]. Pang and Lee, Dumais et al. and Hearst et al. all employ support vector machine (SVM) algorithms in the development of their classifiers [26] [27] [28]. Pang and Lee further extend their study to employ a Naive Bayes classifier- an approach also taken by Lewis [26] [29]. Nigam et al. on the other hand opts for a maximum entropy based classifier which falls under the same category of probabilistic classifiers as Naive Bayes [30]. Other studies such as those conducted by Pang et al. and Boiy et al. take a more holistic approach, by performing a comparison across all these classification approaches, identifying the inherent trade-offs associated with each method [31] [32]. The approach delivers good results, but has the drawbacks of requiring pre-labelled data- which may not always be available- to train the models on, and relying on large volumes of textual data to serve as training, validation and test sets for the algorithms.

2.2.2 Lexicon-Based Approach

The lexicon-based approach is centred on the assumption that the contextual sentiment orientation of a body of text is the sum of the sentiment orientation of each word or phrase in that body of text [33] [34]. The approach can be further divided into two types: a corpus-based approach and dictionary-based approach. The corpus-based approach attempts to find co-occurrence patterns of words to determine the sentiments present in a text [35] whilst maintaining context. The process begins with a "seed list" of opinion words/phrases, and computes sentiment by searching for other opinion-oriented words/phrases in a text which have similar context [1]. Moreno-Ortiz and Fernández-Cruz employ this approach in an effort to integrate domain-specific sentiment analysis into a system initially designed for general language texts. The study achieves promising results, but falls short as the specialization level of the corpus increases while maintaining a static seed list. This is due to the fact that, to an extent, sentiment is lexicalized differently across different areas of specialization [36]. Agarwal and Mittal also report a similar issue with this technique- that it relies only on the polarity of the terms that have appeared in this "seed list", since polarity is computed for the terms that are in the corpus [37].

The dictionary-based approach on the other hand uses a predefined dictionary of words, where each word is associated with a specific sentiment polarity strength [1]. Each text is broken down into single-word tokens- maintaining frequency count but removing context- into what is known as a "bag-of-words" model [7]. Sentiment is then computed by matching words from the text with entries in the employed dictionary [5]. All studies make use of a general language dictionary as a base dictionary for sentiment analysis. While in some cases custom dictionaries are compiled, it is common to employ an already existing and developed lexicon, such as the General Inquirer dictionary- which merges the Harvard IV-4 and Laswell dictionaries to create a comprehensive general English language dictionary [38]. Other such resources include the dictionaries developed by Hu and Liu [39], and Esuli and Sebastiani for the popular SentiWordNet tool [40]. Many studies involving any degree of specialization in their work elect to use a domain-specific glossary in addition to this base dictionary. This is due to the potential for misclassification of terms using a dictionary that is more typical of general language text than domain-specific text such as financial news [5]. Within the domain of finance, Loughran and McDonald address this issue through the development of a finance-specific glossary of terms identified as expressing sentiment in a financial context [41]. Ahmad et al., Zhao et al. and Kelly all employ domain-specific glossaries in addition to general language dictionaries in their studies on the impact of text-based media on financial asset price. They perform analyses on newspapers and other publications in order to investigate how sentiment present within these forms of media influences the value of financial instruments [3] [4] [5] [18] [19] [42].

2.2.3 Existing Sentiment Analysis Systems

As the field of sentiment analysis has developed and gained traction over the years, several attempts have been made at developing proprietary systems capable of performing opinion mining across large datasets. From a research standpoint, Esuli and Sebastiani, Ahmad et al. and Taboada et al. are amongst some of the parties that have developed their own sentiment analysis tools. SentiWordNet is a tool for opinion mining developed by Esuli and Sebastiani that opts for a machine-learning approach to classification. The system is built on top of WordNet- a lexical database for the English language that groups words into sets of synonyms. SentiWordNet associates each set of synonyms to three numerical scores describing how objective, positive and negative the terms contained in the set are. This association is performed by eight ternary classifiers, all characterized by similar accuracy levels but different classification behaviour [40]. Esuli et al. has gone on to develop several iterations of the system, with its current third generation exhibiting significantly stronger results than its predecessors [43]. Ohana and Tierney use the SentiWordNet resource for the purpose of sentiment classification of film reviews, obtaining results that indicated effective performance when compared with manual resources on this specific task, and significant improvement over a pure term counting approach to sentiment analysis [44].

The Sentiment Orientation Calculator (SO-CAL) developed by Taboada et al. uses dictionaries of words annotated with their semantic orientation, and incorporates intensification and negation [45]. Dictionaries can be created manually as done with the General Inquirer dictionary by Stone et al. [38], or automatically using seed words to expand the list of words in the dictionary. Taboada et al. discover that manually built dictionaries generally provide superior results when compared to automatically generated systems.

Rocksteady is an affect analysis system developed at Trinity College Dublin by Ahmad and Zemánková that also employs a dictionary-based approach to sentiment analysis [42]. The system offers a library of both general language dictionaries and domain specific glossaries, allowing the user to select any number of base and specialized dictionaries for the analysis process. The aforementioned studies performed by Ahmad et al., Zhao et al. and Kelly to investigate the impact of text-based media on financial asset price all make use of Rocksteady as an affect analysis system of choice. The finance-oriented nature of these studies prompts the use of a finance domain-specific glossary in addition to a standard general language dictionary [4] [5] [18] [19]. Ahmad et al. also makes use of the Rocksteady tool to analyze news articles in the time leading up to the 2011 Irish elections, in an attempt to predict the outcome. The study yielded promising results that were remarkably close to the final observed reality of the election [6].

It is also worth acknowledging that a wide range of sentiment analysis systems also exist in a commercial capacity. Bloomberg, as part of their "Professional Services" suite of tools, offers sentiment analysis capabilities to customers, believing that sentiment plays a crucial role in the finance industry and has the ability to positively influence trading returns. Infotrie has similarly developed its "FinSentS" product, which analyzes over fifty thousand stocks, topics, people, commodities and other assets, through information gathered from millions of websites, blogs and business news publications in real-time. Lexalytics is another company that has developed its own sentiment analysis system [46], which has formed the foundation for the popular Thomson Reuters News Analytics service used in a plethora of studies such as those conducted by Mitra [47] and Leinweber [48].

2.3 Statistical Modelling

The goal of this study is to examine the effect of sentiment expressed in written media on financial markets- specifically on the values of market indices. As discussed previously, this necessitates a financial data time series, and a corresponding sentiment time series that is obtained through textual sentiment analysis- spanning the same period of time as its financial counterpart. In order to investigate the influence of one time series on the other, it becomes necessary to consider both time series together as a multivariate time series.

In pioneering and highly influential studies, Sims puts forward the use of an econometric technique known as vector autoregression (hereafter referred to as VAR) for the analysis of multivariate time series [49] [50]- a practice that has since been widely adopted by previous studies in this field. VAR models consider multivariate time series as being composed of two types of variables: endogenous variables that are inherently internal in nature to the system, and exogenous variables that are viewed as external to the system. Studies examining the impact of sentiment on financial markets generally consider financial return and sentiment as the endogenous and exogenous variables of the system respectively, as sentiment is regarded as an external factor affecting return. Sims' VAR method performs a multivariate analysis in which the exogenous variables in an equation are used to compute the value of the endogenous variable and its linear dependence on these exogenous variables. Note that each independent variable can also be used as a dependent variable of all other variables [51], thus making it possible to study converse relationships- for example the effect returns has on sentiment. The linear dependence calculation, which is performed using Granger's causality test, serves as an excellent measure of the statistical significance of any present dependencies between the variables that emerge as a result of the analysis [52].

Many macroeconomic econometric studies have since proceeded to employ this method of multivariate time series analysis to investigate linear dependence between variables present in a model. A significant subset of these studies that examine the effect of sentiment in a financial capacity make use of vector autoregressive models in an effort to jointly examine financial and sentiment time series. Baek enlists the technique to study the influence of sentiment on both prices and dividends in the US market, and vice versa, using the S&P 500 index as a proxy for the market [53]. Tetlock and Kelly use VAR models to examine the impact of sentiment on the Dow Jones Industrial Average (DJIA), with Tetlock further breaking down sentiment into the sub-components of pessimism, negative sentiment and weak sentiment [5] [17]. Ahmad et al. takes a unique approach when studying the US market, opting for a more granular approach by using a VAR model in their investigation of the impact of sentiment on firm-level returns rather than on an aggregate level- carefully selecting twenty large US corporations that span across multiple industries to examine [4]. Beyond the US market, Kumari and Mahakud explore the effect of sentiment on both returns and volatility in the emerging Indian market through the use of VAR models- selecting the most popular SENSEX and Nifty indices as market proxies for the study [11]. Finally, Zhao and Ahmad employ a VAR model when studying the effects of sentiment on the Chinese market- in this instance using the SSEC index as a proxy for the market [19].

3 Methodology

This purpose of this chapter is to provide an overview of the steps taken from the commencement of the study, through to its completion. As previously outlined (Chapter 1), the scope of the work undergone is divided into three distinct facets or phases: the acquisition and preprocessing of financial data and media content, sentiment analysis of the media content, and the statistical modelling of the generated financial and sentiment time series. This section is thus appropriately divided into three distinct sections, with each section providing a comprehensive overview of the methodology employed by that phase of the study.

Figure 3.1 provides an overview of the developed system architecture, clearly indicating each of the three phases and the interactions that take place between them in order to deliver the final result: the quantification of any potential relationship that exists between sentiment expressed in formal media, and market returns.

3.1 Data Acquisition and Preprocessing

The data gathered for this study can be clearly divided into two categories: financial time series data and written media content. This section offers an overview of the techniques used to acquire and preprocess this data, and is divided into two sub-sections to reflect the differing approaches taken for each category.

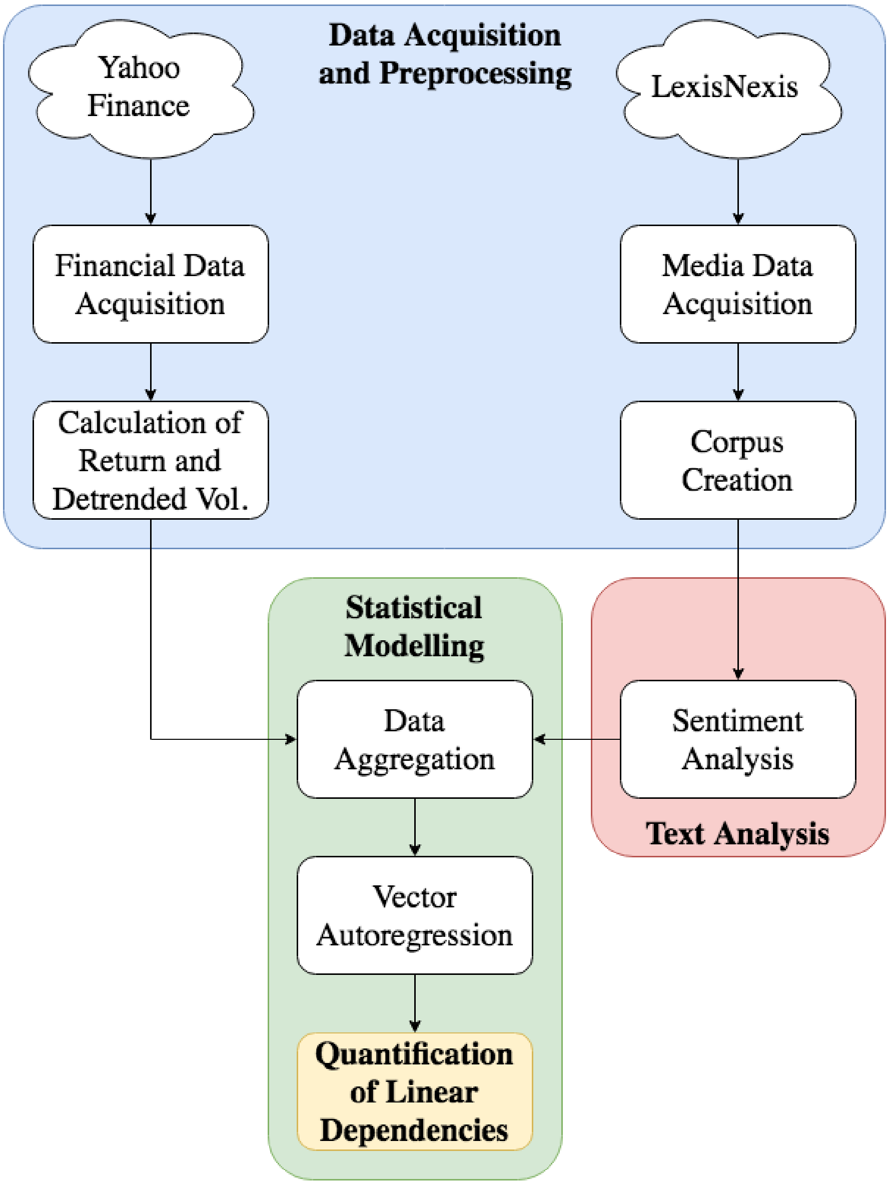

Figure 3.1: System Architecture

3.1.1 Financial Data

As already outlined (Chapter 1), three indices were selected to serve as proxies for each the three markets being examined. Previous studies such as those discussed in the prior chapter (Chapter 2) were consulted in order to identify the most commonly used indices in similar contexts. Table 3.1 shows the markets and corresponding proxy indices that were selected for this study based on this analysis of previous work.

In order to retrieve historical price and volume data for each of these indices, the popular "quantmod" (Quantitative Financial Modelling Framework) package was used in conjunction with the R programming language and software environment- an alternative to the "Quantmod" package used by Kelly [5] that was previously mentioned (Chapter 2).

Table 3.1: Markets and Corresponding Proxy Indices

Market	Selected Proxy Index	Reference Studies
United States	S&P 500	Zhao and Ahmad [18] [19], Fisher and Statman [20]
United Kingdom	FTSE 100	Nikkinen and Vähämaa [21], Gillam et al. [22]
India	SENSEX	Bhardwaj et al. [1], Kumari and Mahakud [11]

Quantmod aggregates financial data from multiple sources, and allows users to select which sources to query for the data. For the purpose of this study, "Yahoo Finance" was the selected source for the data- a well-reputed source for stock and index prices. Price and volume time series data from September 2003 until January 2018 was obtained for each index, and stored as an R-compatible "extensible time series" variable. The start date of this time period of interest was dictated by the date that the SENSEX index first shifted to the free-float market capitalization methodology. The index was the last of the selected indices to employ this technique- doing so on September 1st 2003 [54] [55]. The new methodology offered an alternative way of calculating the market capitalization of an index's constituent companies [56]. By ensuring all indices to be examined used the same methodology, it became possible to perform a fair analysis across all markets.

Upon examination of the historical price data obtained, it was identified that prices were highly correlated- i.e. the price at any given time was influenced by previous prices- thus presenting a challenge of constructing a fair model due to the presence of historical trends in the data. In order to overcome this problem, a return time series was calculated in R for each of the three indices, using the formula shown in Equation 1:

$$r_t = \log_{10}\left(\frac{p_t}{p_{t-1}}\right) \tag{1}$$

where p_t is the price at time t. Return is used as the dependent variable for statistical modelling by the vast majority of previous studies in lieu of price due to the absence of a correlation between its values. It was further identified that the obtained trade volume time series carried with it undesirable historical trends. These trends were removed through the use of the sixty-day rolling average of log volume (Equation 2), rather than trade volume itself- a detrending methodology employed by Tetlock [17], originally developed by Campbell et al. in 1993 [57].

$$V'_t = \frac{1}{60}\sum_{i=t-60}^{t-1} \log_{10} V_i \tag{2}$$

where V_i is the volume at time i respectively.

3.1.2 Media Content

To facilitate the extraction of emotion and opinions expressed towards a financial market from media publications, it was necessary to build a corpus of news articles specific to each market. The study focuses specifically on media content published by formal media sources, making it necessary to identify an appropriate sample set of sources. A plethora of news sources exist that span across the three markets being investigated, and hence careful consideration was made to ensure the selected sample served as an accurate representation of the formal media available in that market. Two types of sources were selected: top-selling and highly-circulated general publications, and finance-specific publications. These were divided between physical print and web-based publications. Table 3.2 provides an overview of the formal media publications that were selected as sources for this study.

Table 3.2: Market-Specific Formal Media Publications

Market	Publication	Type
United States	Wall Street Journal[2]	web
	CNN	web
	MSNBC	web
	USA Today	print
	The New York Times	print
	The New York Post	print
	The Los Angeles Times	print
	Daily News	print
	The Washington Post	both
United Kingdom	The Financial Times	web
	The Telegraph	web
	The Express	print
	The Times	print
	The Mirror	print
	The Daily Star	print
	The Daily Mail & Mail on Sunday	print
	The Sun	print
	The Daily Record & Sunday Mail	print
	The Guardian	both
India	The Times of India	web
	The Economic Times	web
	The Hindustan Times	print
	The Hindu	print
	The Hindustan	print
	The Indian Express	print

[2] Using the Lexis Nexis News and Business service, it was only possible to obtain the abstracts of the Wall Street Journal articles

Given the wide range and diversity of the sources selected, the study opted to use the Lexis Nexis News and Business service to obtain the large volume of news articles, as similarly done by Kelly, Zhao and Ahmad [5] [18] [19]. The service provided the functionality to perform a search across all published and available articles from pre-selected sources using one or more keywords. The keyword used for each search query was the name of the index being used as a proxy for the market in question. Lexis Nexis also offered the option to specify search criteria, which refined the search by searching for the query term(s) only in certain contexts. The list of available criteria included: anywhere in the text, in the headline, at the start[3], in the company name, within the indexing terms, major mentions[4], three or more mentions, and in the byline. Table 3.3 summarizes the query terms and criteria employed by this study for use with the Lexis Nexis service.

Table 3.3: Lexis Nexis Query Terms and Criteria

Market	Query Term	Query Criteria
United States United Kingdom India	"s&p 50" "ftse 100" "sensex"	Major mentions[4], 3 or more mentions

When compiling the corpora specific to each market, articles containing the query term (the name of the proxy index specific to that market) were returned if that query term had any major mentions[4], or occurred in the text three or more times. This range of criteria was selected under the belief that the corpus created would contain a good balance of both size and relevance. The use of more restrictive criteria (for example, searching for the query term only in the headline) was also experimented, which yielded a higher degree of relevance in the articles present in the corpus. However, this also resulted in a significant decrease in the size of the corpus, which was not conducive to effective analysis over such a large time period of interest.

Due to service restrictions, it was only possible to obtain a maximum of five hundred articles at any given time from the Lexis Nexis tool. Hence in order to obtain all the articles covered by the fifteen-year time period, it was necessary to break the search down over smaller periods, to ensure that all articles for the period of interest were retrieved. The text files retrieved from the sub-searches were then concatenated, thus creating a single corpus file containing all the news articles, from all selected sources, for each market.

[3] Refers to mentions in the headline or lead paragraph

[4] Refers to mentions in the headline, lead paragraph or indexing terms

3.2 Text Analysis

The goal of the text analysis phase was to identify, quantify and extract any sentiment present in each of the three corpora generated in the previous phase. Following much of the literature in this field, such as the work of Ahmad et al. [4], Kelly [5] and Tetlock [17] to name a few, a focus is placed specifically on negative sentiment present in media. As previously discussed (Chapter 2), there are several methods to sentiment analysis that may be employed. This study opted for a lexicon-based approach in conjunction with a "bag-of-words" text representation model. The advantage of this approach is that standard lexicons such as the those mentioned previously (Chapter 2) have been created and extensively developed by many influential studies, covering sentiment across a broad range of contexts- and these resources have been made publicly available for use.

3.2.1 Rocksteady Affect Analysis System

For the purpose of sentiment analysis, the aforementioned Rocksteady affect analysis system (Chapter 2) was employed. Developed at Trinity College Dublin by Ahmad and Zemánková [42], and used by Ahmad et al., Zhao et al. and Kelly [4] [5] [6] [18] [19], the program analyzes a user-provided corpus of texts and quantifies the level of sentiment present across it in time series form. Flexibility is provided to the user through functionality such as the ability to group articles in different ways (e.g. by source, by day, by month, etc.), exclude duplicate articles, and perform search queries on the corpus as a whole. For ease of data aggregation in the statistical modelling phase, articles were grouped by day, thus creating a sentiment time series matching the daily frequency of the already obtained financial time series.

Rocksteady uses a "bag-of-words" model [7] to represent text, whereby an article is deconstructed into a "bag" of its words, removing context while maintaining word frequencies. Word tokens were are matched against a predetermined dictionary of affect-laden words as defined by authoritative sources in order to compute the sentiment present in any given text. Table 3.4 provides an example of how the system uses this technique to quantify the level of negative sentiment present in a sample text. First, the text is broken down into single word tokens, keeping a track of their frequency (Table 3.4a). By matching these tokens with a glossary of terms carrying negative connotations, a simple calculation is then performed in order to determine the percentage of the text that is deemed to exhibit negative sentiment (Table 3.4b).

Table 3.4: Example of Sentiment Analysis using Rocksteady to extract Negative Sentiment

Sample Text: *"Sensex dips sharply, down 91 points."*

(a) Bag-of-Words Representation

Term	Frequency
Sensex	1
dips	1
sharply	1
down	1
91	1
points	1
TOTAL	**6**

(b) Calculation of Negative Sentiment Score

Glossary hits for Negative Sentiment	2
Total Number of Terms	6
Negative Sentiment % Score	**33.33%**

Note: sample text is an excerpt from a June 3rd 2009 article in the Hindustan Times.

Rocksteady facilitates the use of multiple dictionaries divided into two categories: base dictionaries and specialist dictionaries. Base dictionaries consist of general terms that are used in multiple contexts, while specialist glossaries are comprised of domain-specific terms. Any terms present in both base and specialist dictionaries take on the attributes defined in the specialist glossary. A general English language dictionary was selected as the base dictionary, coupled with a finance-specific glossary in order to avoid the potential misclassification of terms in a financial context as discussed by Loughran and McDonald [41]. Rocksteady provides an integrated oil and finance glossary, however after its examination it was discovered that the glossary did not provide an exhaustive list of finance-specific terms or distinguish between finance-specific and trading-specific terms. For these reasons, it was decided to create a new glossary using the existing one as a foundation, and extend it in an effort to rectify these issues and hence improve the accuracy of analysis.

3.2.2 Domain-Specific Glossary Creation

The base dictionary employed for the study included a comprehensive list of terms associated with the expression of sentiment. The main purpose of the domain-specific glossary was to ensure that certain terms and phrases normally associated with sentiment in general language- that take on different connotations in the area of this study- were not misclassified in the sentiment analysis process. For example, the word "close" in a normal context is often noted as expressing negative sentiment in a general language context. For example, the phrase "the shop will close down" is regarded in general as being a negative phrase. However in the world of trading, the "close" price of a stock is simply the price of the stock at the end of a trading day, and hence there is no sentiment attached to the word.

Alternatively in business, to “close a deal” refers to agreeing upon the terms and conditions of a deal, and accepting them as satisfactory. In certain contexts, this can be construed as taking on a positive connotation. This is one example of a term that was omitted from the domain-specific glossary native to Rocksteady.

In an effort to combat this issue in the most effective manner possible, three domains were selected as being of crucial importance to this study: the domains of finance, economics and trading. For this reason, the new glossary focuses primarily on the inclusion of terms specific to these three domains. To select these terms, it was decided to include those deemed relevant by authoritative sources in each of the areas. Specifically within the field of economics, The Economist is a well-known and highly-regarded source that has published a glossary of key economic terms [58]- adapted from “Essential Economics” by Matthew Bishop [59]- each of which was included in the new specialist glossary. The Government of Australia's Department of Industry, Innovation and Science has similarly published a list of key financial terms [60], all of which were also included in the new specialist glossary. Lastly, India Infoline- a leading integrated financial services group in India [61]- has published a glossary of terms commonly associated with the stock market [62]. Each of these terms was included as a trading-specific term in the new specialist glossary. Table 3.5 shows a summary of these selected sources and the respective domains that their glossaries cover. The authoritative and well-reputed nature of these sources justifies their consultation for the purposes of constructing a credible and accurate glossary.

Table 3.5: Domain-Specific Glossary Term Sources

Domain	**Source**
multiple	Rocksteady Integrated Oil, Finance and Economics Glossary
Economics	The Economist [58]
Finance	Department of Industry, Innovation and Science, Government of Australia [60]
Trading	India Infoline [62]

3.3 Statistical Modelling

The statistical modelling phase of work served as a bridge between the work completed in the prior two phases- aggregating the financial time series data obtained, with the proxy for negative sentiment present in formal media that was extracted for each market using sentiment analysis. A technique from econometrics known as vector autoregression was then employed to analyze the resulting multivariate time series, modelling return as a function of a number of influencing independent variables including sentiment.

3.3.1 Data Aggregation

Both a financial return time series and a sentiment time series were extracted from the analysis of the collected data for each market, through methods outlined previously. Given that both time series spanned the same period of time, and both were taken at a daily frequency, it became possible to combine them into a multivariate time series. This was done using R, joining both time series by matching daily values for return and sentiment. A lookup function was written to match the dates across both time series and pair each return value with the corresponding sentiment value for that day, thus resulting in the construction of a multivariate time series.

3.3.2 Accounting for Recession and Human Behaviour

While the aim of the study is to investigate the impact of sentiment on financial markets (specifically, returns), it is naive to assume that sentiment is the only influencing independent variable. Sudden periods of economic downturn, and elements of human behaviour are both factors that potentially hold an influence over returns. This study provisions for these periods of economic downturn, and addresses human behaviour to a limited extent, through the addition of two more variables to the aforementioned multivariate time series. A variable named "*Rec*" was introduced in order to account for recessionary periods, which took on the following properties:

$$Rec_t = \begin{cases} 1, & \textit{if } \text{t is a date within a period of recession} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

It is also suggested that periods following major holidays play a role in the deviation of prices from the intrinsic values of their respective assets, which in turn affects returns. This study focuses on the period following the major Christmas and New Year holidays, by introducing a variable "*Jan*" that takes on the following properties:

$$Jan_t = \begin{cases} 1, & \textit{if } \text{t is a date within the month of January} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Previous studies such as that conducted by Damodaran suggest that there is also a "weekend effect" that must be accounted for, whereby the two working days either side of the weekend (Friday and Monday) exhibit unusual price movements [63]. However, this has since been further examined- such as in the study performed by Alt et al.- to reveal that while this trend is present in historic data, the statistical significance of this weekend effect has since diminished and "disappeared" in recent times [64]. These examples of influencing factors are what are known as "calendar effects", and their importance for consideration is outlined by Taylor in his book titled "Asset price dynamics volatility and prediction" [65], and by Ahmad et al. [4] [51].

3.3.3 Vector Autoregression (VAR) Analysis

For the analysis of the generated multivariate time series, vector autoregressive techniques put forward by Sims as previously discussed (Chapter 2) were used. Return was selected as the endogenous variable of the system, where it served as a proxy for financial market performance. The variables created to account for calendar effects ("*Rec*" and "*Jan*") were included as two of the exogenous variables of the system. Negative sentiment was modelled in two different ways, through the use of two different proxies both considered as independent variables: article count as a preliminary proxy, and the "*Neg*" time series generated by Rocksteady after sentiment analysis. It was expected that while a valid proxy, article count would be found contribute a lesser impact on returns, as discovered by Ahmad et al. [4]- thus validating the use of sentiment analysis in extracting a stronger proxy for negative sentiment. Finally, the detrended trading volume calculated during the preprocessing phase of the study was also included as an exogenous variable in accordance with several previous studies [5] [17] [57].

VAR models also allow for the inclusion of lagged values of both the dependent and independent variables, in order to determine if past values of a particular variable have an influence on its current value. Five lags were selected as the number of lags to factor into the VAR analysis, due to their being five days in the trading week. Rather than simply constructing a single model including all independent variables, a panel approach was taken, whereby a first model was constructed without any exogenous variables, and successive models each built on the previous by introducing additional exogenous variables. This method aided in revealing the effectiveness of considering each exogenous variable, through the examination of each model's residual term and "goodness of fit" (R^2) term. If found that the introduction of a new exogenous variable to the model resulted in an increase in the R^2 value and decrease in the residual term, it would indicate that the new model was a better fit for the data set with a higher accuracy. Thus it would be shown that accounting for that newly-introduced variable was justified in the modelling of return. Note that in the event of any autocorrelation present in returns due to the heteroskedastic nature of the data across all three markets, Newey-West standard errors were introduced to the models [66] [67] in order to assist in its removal.

To perform the VAR analyses, the GNU Regression, Econometrics and time series Library (hereafter referred to as GRETL), was used- an open-source statistical package commonly used for econometric analysis. The program has both a graphical user interface (GUI) and command-line interface, providing users with a wide range of resources and functionalities such as, but not limited to, a variety of estimators, time series methods and limited dependent variables. The primary interest in GRETL was for its multivariate time series VAR analysis tool, which facilitated the analysis of the constructed multivariate time series data for each market. Given a particular model comprising of endogenous and exogenous variables, GRETL's VAR tool performs calculations to determine the coefficients of the exogenous variables. It additionally employs Granger's causality test to evaluate linear dependence of the endogenous variable on these variables, thus providing an indication on the statistical significance of any present dependencies [52].

The generalized VAR model for return as a function of its lagged terms is shown below (Equation 5):

$$r_t = const. + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + ... + \alpha_n r_{t-n} + \epsilon_t \tag{5}$$

where r_t and ϵ_t are the return and residual terms at time t respectively, and α_i is the weight of the returns over the previous i periods. As mentioned previously, it was decided to account for five lags of return, which lead to the first model used for VAR analysis in GRETL (Equation 6):

$$r_t = const. + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \alpha_3 r_{t-3} + \alpha_4 r_{t-4} + \alpha_5 r_{t-5} + \epsilon_t \tag{6}$$

This equation can be simplified by representing all of the lag terms by a single expression as follows (Model 1):

$$r_t = const. + \alpha L_5 r_t + \epsilon_t \tag{Model 1}$$

The second VAR model built on the first by introducing the proxy variables for the calendar effects discussed previously- specifically to account for abnormalities in return during periods directly after major holidays and periods of recession. The new model was then represented as shown below (Model 2):

$$r_t = const. + \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \epsilon'_t \tag{Model 2}$$

where β and γ were the weight coefficients of the proxy variables representing periods directly after major holidays, and periods of recession respectively.

VAR analyses were first performed on these two simple models prior to investigating more comprehensive models that include additional independent variables, such as detrended volume, media article count or the negative sentiment proxy obtained from sentiment analysis of the corpus specific to each market. This was done in order to determine whether or not the proxy variables for calendar effects hold any statistical significance, prior to proceeding with the other models that all include these variables. After running the model through GRETL to obtain results, the R^2 and residual sum squared (RSS) terms were examined. In the event that the following properties were found to hold (Equations 7 and 8), and the coefficients of the calendar effect variables were proven to be statistically significant by Granger's causality test, the study will have justified the consideration of periods following major holidays and periods of recession as influencing factors on return, through the production of a better fitting, higher-degree accuracy model of return.

$$R^2_{Model2} > R^2_{Model1} \tag{7}$$

$$RSS(\epsilon'_t) < RSS(\epsilon_t) \tag{8}$$

The residual term from the first model can also be expressed in terms of the new residual term and proxy variables as follows (Equation 9):

$$\epsilon_t = \beta Jan_t + \gamma Rec_t + \epsilon'_t \tag{9}$$

The third VAR model introduced five lags of the detrended trading volume as exogenous variables in the model, to investigate any potential influence that they may have on return. This new model was represented as shown below (Model 3):

$$r_t = \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \delta Vol_t + \epsilon''_t \tag{Model 3}$$

where δ was the weight coefficient of the variable representing detrended trading volume. The fourth VAR model added the article count proxy variable for negative sentiment into the model. The aim of this model is to investigate whether the findings of Ahmad et al. hold true, in that media article count is a viable proxy for negative sentiment, albeit not the strongest when compared to a counterpart employing sentiment analysis techniques [4]. Note that this model was divided into two sub-models, one including detrended volume (Model 4a) and one not (Model 4b), in order to examine whether the inclusion of detrended volume had any impact on the potential statistical significance of article count. These models are shown below (Model 4a and Model 4b):

$$r_t = \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \delta Vol_t + \lambda_{count} NumArticles_t + \epsilon'''_t \tag{Model 4a}$$

$$r_t = \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \lambda_{count} NumArticles_t + \epsilon'''_t \tag{Model 4b}$$

where λ_{count} was the weight coefficient of the proxy variable representing article count. The fifth and final VAR model introduced the negative sentiment proxy variable generated by Rocksteady from its analysis of the compiled corpus for each market. Serving as an alternative to article count, this variable was expected to serve as a stronger and more accurate proxy variable for negative sentiment than article count. Note that again this model was divided into two sub-models, one including detrended volume (Model 5a) and one not (Model 5b), in order to examine whether the inclusion of detrended volume had any impact on the potential statistical significance of the negative sentiment proxy. These models are shown as follows (Model 5a and Model 5b):

$$r_t = \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \delta Vol_t + \lambda_{neg} Neg_t + \epsilon''''_t \tag{Model 5a}$$

$$r_t = \alpha L_5 r_t + \beta Jan_t + \gamma Rec_t + \lambda_{neg} Neg_t + \epsilon''''_t \tag{Model 5b}$$

where λ_{Neg} was the weight coefficient of the proxy variable representing negative sentiment. After running each of these models through GRETL to obtain results, the R^2 and residual sum squared (RSS) terms were examined. In the event that the following properties were found to hold (Equations 10 and 11), and the coefficients of each introduced independent variable were proven to be statistically significant by Granger's causality test, the study would have justified the consideration of each of these variables (detrended volume, media article count and the negative sentiment proxy obtained from Rocksteady) as influencing factors on return, through the production of continuously better fitting, higher-degree accuracy models of return.

$$R^2_{Model5} > R^2_{Model4} > R^2_{Model3} > R^2_{Model2} > R^2_{Model1} \tag{10}$$

$$RSS(\epsilon''''_t) < RSS(\epsilon'''_t) < RSS(\epsilon''_t) < RSS(\epsilon'_t) < RSS(\epsilon_t) \tag{11}$$

The original residual term from the first model can again be expressed in terms of the new residual term and all four exogenous variables as follows (Equation 12):

$$\epsilon_t = \beta Jan_t + \gamma Rec_t + \delta Vol_t + +\lambda_{neg} NumArticles_t + \epsilon''''_t \tag{12}$$

If the results of each successive VAR analysis indicated the statistical significance of the newly introduced exogenous variable, and the R^2 and residual terms were observed to continuously increase and decrease respectively, it can be claimed that each model had extracted some meaning from the original residual term, thus improving the fit of the model and hence improving its accuracy.

4 Results and Discussion

This chapter provides a detailed overview and discussion of the results obtained from each phase of work, using the methodologies outlined in the previous chapter (Chapter 3). The statistical models developed for the three financial markets under study are used to model return- both over the full time period encapsulated by the study, and over the sub-periods before, during and after the “Great Recession”- with a comparison and contrast between the obtained results provided.

4.1 Data Acquisition and Preprocessing

This section offers an overview of the gathered financial time series data and media content, and is divided into two sub-sections to reflect the differing approaches to retrieval taken for each category.

4.1.1 Financial Data

The financial data was obtained using the “quantmod” (Quantitative Financial Modelling Framework) package in R, specifying Yahoo Finance as the data source. Index price and volume was obtained for each of the markets from 1st September 2003 until 18th January 2018. As discussed in Chapter 3, the correlation present in the price data was removed through the use of log returns (Equation 1). Additionally- following previous studies such as those performed by Kelly and Tetlock- the volume time series was detrended through the calculation of a sixty-day moving average of volume [5] [17] (Equation 2). Figure 4.1 shows the return calculated for each of the market indices, in the form of their density curves.

The distribution of return data for each market bears resemblance to that of a normal distribution in that the data is relatively symmetric about the mean, however it is observed that each of the curves has a “sharper peak” and “fatter” tails than when compared to a normal distribution, indicating a wider spread of returns than what would be described by the standard bell-curve. Thus, aligned with the findings of studies such as those performed

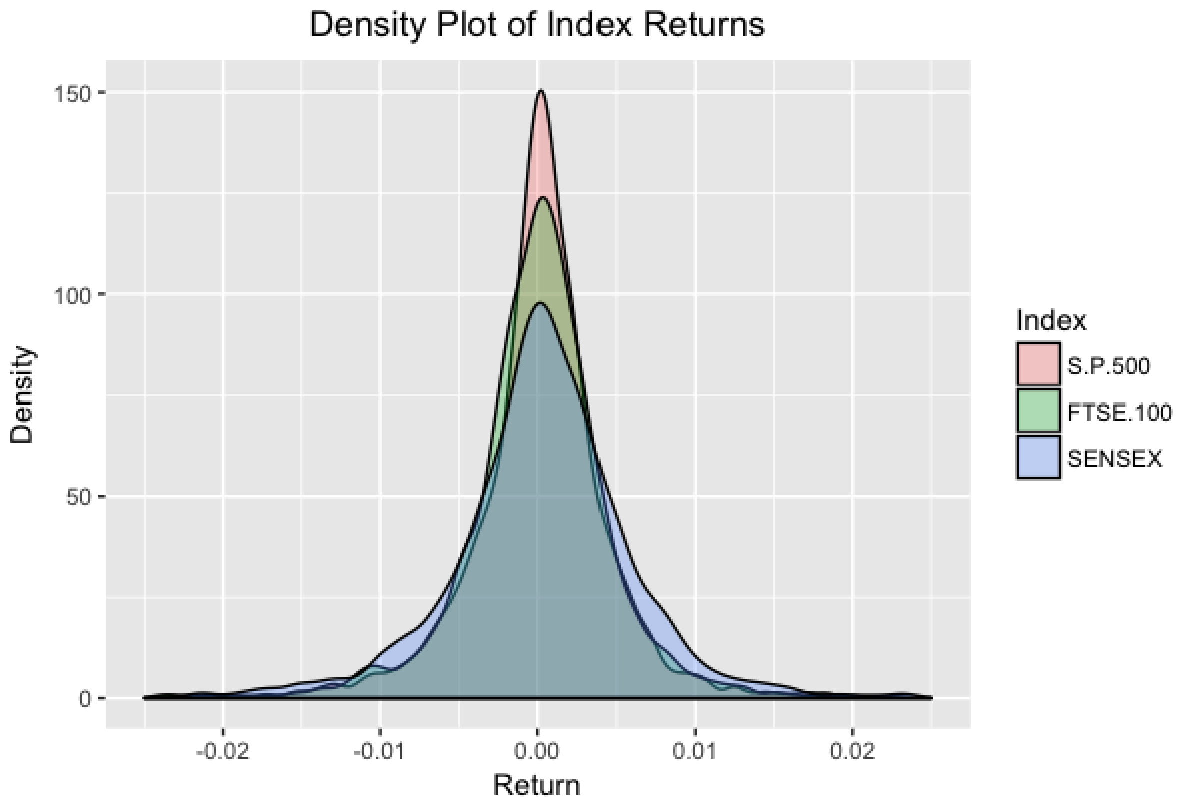

Figure 4.1: Density Curves of Index Returns

by Brown and Warner, Andersen et al. and Fama, the common traditional assumption that returns are normally distributed is shown to be untrue for this particular data set [9] [68] [69]. These claims are supported by the descriptive statistics of the density curves shown in Table 4.1. The returns across all three markets exhibit significantly higher kurtosis values- 12.55, 8.82 and 9.63 for the US, UK and Indian markets respectively- than that expected of a normal distribution ($k_{normal} = 3$) and slight negative skewness values.

It is worth noting that these descriptive statistics are calculated across a relatively small number of observations (approximately 3,500), and hence the figures for skewness and kurtosis are used primarily as a suggestion of non-normality. To further support these claims, the relative and average frequencies for the samples of returns are shown in Tables 4.2 and 4.3, taking care to include baseline values for the normal distribution. These results clearly indicate that the return for each market's proxy index exhibits a stronger concentration of values towards the centre of the distribution, and a wider spread of return than that of a normal distribution, giving the distinct "sharper" distribution shape with "fatter" tails.

The time period of interest is an interesting choice for study, due to the significant economic recession that occurred within that period, between December 2007 and June 2009 [8] as defined by the US National Bureau of Economic Research (NBER). For this reason, the study further elects to perform analyses across three subsets of data for each market-

Table 4.1: Descriptive Statistics of Return

		USA	**UK**	**India**
Mean	[E-05]	12.27	6.77	23.85
Standard Error	[E-05]	8.32	8.07	10.63
Median	[E-04]	2.95	2.12	3.98
Mode	-	0	0	n/a
Standard Deviation	[E-03]	5.01	4.86	6.30
Sample Variance	[E-05]	2.51	2.36	3.97
Kurtosis	-	12.55	8.82	9.63
Skewness	-	-0.36	-0.16	-0.11
Range	[E-02]	8.87	8.10	12.07
Minimum	[E-02]	-4.11	-4.02	-5.13
Maximum	[E-02]	4.76	4.08	6.94
Sum	-	0.44	0.25	0.84
Count	-	3620	3622	3511
Confidence Level (95.0%)	[E-04]	1.63	1.58	2.08

[E-XX] indicates the power to which the stated number is raised.

Table 4.2: Relative Frequencies for Samples of Return

	no change	**Within**		**Beyond**						
		0.25	*0.5*	*1*	*1.5*	*2*	*3*	*4*	*5*	*6*
Normal	-	19.74%	38.29%	31.73%	13.36%	4.55%	0.27%	0.01%	0.00%	0.00%
Series										
S&P 500	0.05%	34.77%	57.26%	19.12%	8.54%	4.92%	1.81%	0.82%	0.44%	0.22%
FTSE 100	0.14%	28.87%	52.73%	20.39%	9.76%	5.34%	1.44%	0.76%	0.30%	0.19%
SENSEX	0.00%	29.47%	51.94%	21.69%	9.92%	5.26%	1.51%	0.64%	0.22%	0.08%
Averages										
All Series	-	31.55%	53.91%	21.09%	9.77%	5.23%	1.63%	0.75%	0.28%	0.12%

italicized numbers represent the number of standard deviations from the mean.

Table 4.3: Average Frequencies for Standardized Daily Return

Range	Observed	Normal	Diff.
0 - 0.25	31.55%	19.74%	11.81%
0.25 - 0.5	22.36%	18.55%	3.81%
0.5 - 1	25.00%	29.98%	-4.98%
1 - 1.5	11.32%	18.37%	-7.05%
1.5 - 2	4.54%	8.81%	-4.27%
2 - 3	3.60%	4.28%	-0.68%
3+	1.63%	0.27%	1.36%
Total	100.00%	100.00%	0.00%

italicized numbers represent the number of standard deviations from the mean.

specifically the time periods before, during and after the recessionary period. Table 4.4 shows a breakdown of the key descriptive statistics for each market across these three periods, expressed relative to their values in the period before recession. It can clearly be seen that there is a considerable fluctuation in the descriptive statistics across the three sub-periods examined. The dramatic change in the variance across each market alludes to the heteroskedastic nature of financial return across all three markets.

Table 4.4: Breakdown of Descriptive Statistics by Recessionary and Non-Recessionary Periods

		Mean	**Variance**	**Kurtosis**	**Skewness**
USA	before	1	1	1	1
	during	-3.34	10.23	1.96	0.13
	after	1.44	1.54	2.59	1.71
UK	before	1	1	1	1
	during	-2.87	7.50	1.33	-0.18
	after	0.76	1.53	0.86	0.44
India	before	1	1	1	1
	during	-0.71	3.53	0.53	-0.45
	after	0.30	0.48	0.34	0.29

Note the "before" period is used as a baseline.
The recessionary period occurred between 12/2007 and 06/2009 according to the National Bureau of Economic Research [8].

4.1.2 Media Content

The retrieval of news articles for each market's corpus of texts to be constructed was performed by making use of the Lexis Nexis News and Business aggregation service. Both print and web-based publications with high circulation and readership were selected as sources for the media content. The query term used to gather texts pertaining to each of the three markets was the name of the index employed as a proxy for that market. Additionally, the query criteria was restricted to return articles with major mentions, or three or more mentions of the query term (recall Table 3.3).

Table 4.5 shows a breakdown of the articles obtained from each publication across the three markets. Note that the volume of articles obtained for the UK market is significantly higher than for the US or Indian markets. This is perhaps partly attributed to the fact that the FTSE 100 index originated as a joint venture between the London Stock Exchange and The Financial Times- which happens to be one of the media sources selected for the UK market. Thus the Financial Times frequently reports on the performance of the index, which is clearly indicated in the table by the elevated number of articles attributed to the single Financial Times source.

Table 4.5: Breakdown of Articles Retrieved by Publication

USA		UK		India	
Name	Count	Name	Count	Name	Count
The New York Times (p)	1103	The Financial Times (w)	8638	The Times of India (w)	2368
Wall Street Journal[5](w)	1411	The Times (p)	8621	The Economic Times (w)	2227
USA Today (p)	773	The Telegraph (w)	7214	The Hindu (p)	406
The New York Post (p)	158	The Guardian (b)	5708	The Hindustan Times (p)	2245
The Washington Post (b)	836	The Sun (p)	4397	The Hindustan (p)	14
The Los Angeles Times (p)	19	The Mirror (p)	353	The Indian Express (p)	1070
Daily News (p)	145	The Express (p)	1911		
CNN (w)	11	The Daily Star (p)	1042		
MSNBC (w)	15	The Daily Mail & Mail on Sunday (p)	1752		
		The Daily Record & Sunday Mail (p)	256		
TOTAL	**4471**	**TOTAL**	**39892**	**TOTAL**	**8330**

(p) print source (w) web-based source (b) both

*Using the Lexis Nexis News and Business service, it was only possible to obtain the abstracts of the Wall Street Journal articles

4.2 Text Analysis

4.2.1 Sentiment Analysis using the Rocksteady Engine

The Rocksteady system was used in order to perform sentiment analysis on each of the market-specific corpora created. This was done using a dictionary-based approach to sentiment classification, in conjunction with a base English dictionary and an appropriate custom domain-specific glossary. Articles were grouped by day, and a proxy time series for negative sentiment was obtained for each market, which are shown in Figure 4.2[6].

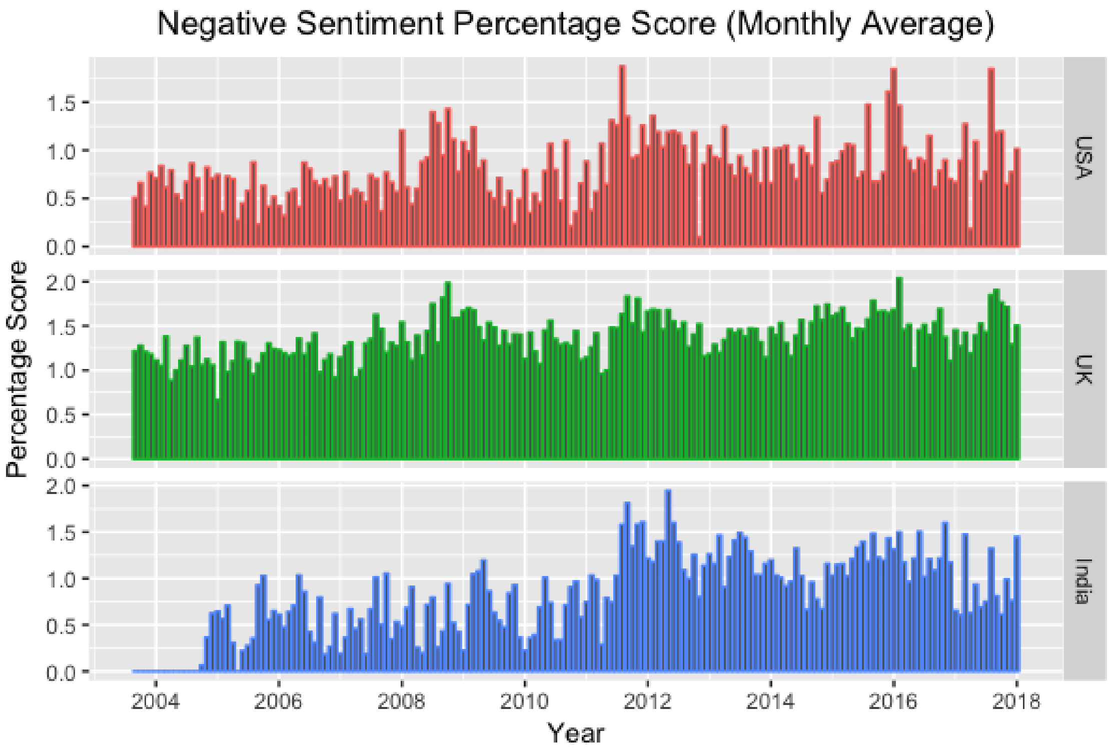

Figure 4.2: Negative Sentiment Proxy Time Series[6]

As previously discussed (Chapter 3), prior to the statistical modelling of return accounting for this negative sentiment proxy, a preliminary model was constructed examining the potential effect of article count on market performance. Hence, a time series for article count was also obtained for each market at a daily frequency from the Rocksteady system, which is shown in Figure 4.3[6].

[6] Time series observations were taken at a daily frequency. However for the purpose of visualization in this document, the monthly average is shown.

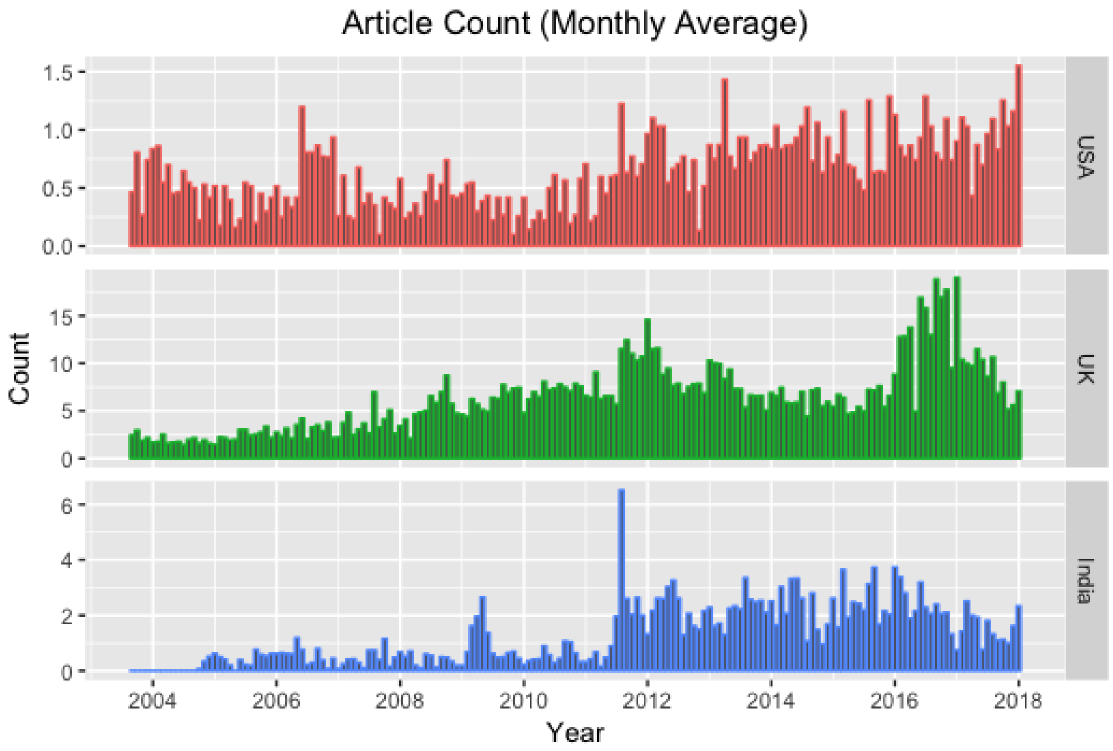

Figure 4.3: Article Count Time Series[6]

4.2.2 Domain-Specific Glossary Creation

The domain-specific glossary created for use in conjunction with the base English dictionary provided by Rocksteady, comprised of terms specific to the fields of economics, finance and trading, as dictated by authoritative sources. The "Integrated Oil, Finance and Economics" glossary provided by Rocksteady was used as a foundation, in the anticipation that its expansion would further reduce the potential for the misclassification of terms in the context of this study. Table 4.6 provides a breakdown of the number of terms added to the new glossary, that were obtained across each of the consulted sources.

Table 4.6: Custom Domain-Specific Glossary

Domain	Source	Terms	New Terms	Expanded Terms[7]
multiple	Integrated Oil, Finance and Economics Glossary	4821	n/a	n/a
Finance	Department of Industry, Innovation and Science, Government of Australia [60]	144	142	2
Economics	The Economist [58]	703	547	156
Trading	India Infoline [62]	165	136	29
	TOTAL	**5833**	**825**	**187**

4.3 Statistical Modelling

This statistical modelling of return accounting for sentiment is divided into two principal case studies: an analysis over the full time period examined by the study, and an analysis across recessionary and non-recessionary periods present within that period. Note that in all of the following VAR analyses, Newey-West standard errors were introduced to the models in order to assist in the removal of any autocorrelation present in returns due to the heteroskedastic nature of the data across all the markets. While this showed to be effective, the final results were still shown to exhibit signs of autocorrelation in returns, although it should be noted that their impact was significantly lessened.

4.3.1 Case Study: Full Time Period

VAR analyses were first performed on models 1 and 2 across each of the three markets, in order to investigate whether or not the introduction of the proxy variables for calendar effects alone were shown to carry any statistical significance. The results of this analysis across each of the three markets are shown in Table 4.7.

It is observed that proxy variable accounting for economic recession carries statistical significance in the US market, negatively contributing 8.84 basis points towards returns, however does not seem carry any statistical significance in the UK or Indian markets. Furthermore, the proxy variable accounting for the period following the major Christmas and New Year holiday is shown to bear no statistical significance for this time period across all markets. Note that the residual sum squared (RSS) values for each market is shown to decrease slightly upon the introduction of the calendar variables, and in turn the R^2 is shown to rise considerably, indicating that Model 2 is in fact a better fitting model than Model 1 which just accounted for lagged values of return.

For each market, additional panel regressions were then performed, accounting for the other computed independent variables believed to have an influence on return. Model 3 introduced detrended volume to the model in addition to the calendar variables. Prior to examining the impact of negative sentiment on returns, a preliminary analysis was performed to investigate whether or not article count had an influence, and this is reflected in Model 4. Finally Model 5 introduces the proxy for negative sentiment computed using the Rocksteady system. Note that both Models 4 and 5 are divided into two sub-models each, including and excluding detrended volume respectively. Tables 4.8, 4.9 and 4.10 show the results of these panel VAR analyses across the US, UK and Indian markets.

[7] Expanded terms refer to terms that already existed in the integrated glossary used by Rocksteady, but who's affect category definitions were updated.

Table 4.7: Panel Regressions- Models 1 and 2 (01/09/2003 - 18/01/2018)

	USA		**UK**		**India**	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
	$n = 3555$	$n = 3555$	$n = 3555$	$n = 3555$	$n = 3446$	$n = 3446$
Constant	1.45	2.73***	0.74	1.78**	2.31**	3.61***
r_{t-1}	-1049.20***	-1082.14***	-380.55	-401.18	734.60**	716.50**
r_{t-2}	-654.08	-689.10	-502.88	-523.14	-543.56*	-559.26*
r_{t-3}	159.01	122.06	-456.26	-475.38	-175.40	-192.12
r_{t-4}	-270.82	-305.54	342.17	323.51	-95.67	-112.02
r_{t-5}	-456.01	-489.08	-589.74*	-606.22*	-326.55	-343.83
Cal. 1: *Jan*	-	-3.40	-	-3.33	-	-5.43
Cal. 2: *Rec*	-	-8.84*	-	-6.67	-	-7.48
RSS	0.0884	0.0881	0.0839	0.0837	0.1354	0.1351
R^2	0.0173	0.0207	0.0105	0.0127	0.0095	0.0115

Coefficient terms shown in basis points (1 BP = 0.01%)
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 4.8: USA Panel Regression- Models 3, 4 and 5 (01/09/2003 - 18/01/2018)

	Model 3	**Model 4a**	**Model 4b**	**Model 5a**	**Model 5b**
			$n = 3555$		
Constant	-48.68	-50.94	3.39**	-46.09	2.09
r_{t-1}	-1080.96***	-1070.67***	-1070.72***	-1039.32***	-1039.34***
r_{t-2}	-691.75	-689.40	-686.72	-684.96	-680.77
r_{t-3}	108.94	108.34	120.84	152.39	167.16
r_{t-4}	-300.26	-308.14	-312.93	-318.06	-320.77
r_{t-5}	-485.66	-482.13	-485.05	-472.24	-473.45
$ArticleCount_t$	-	0.89	0.90	-	-
$ArticleCount_{t-1}$	-	-0.98	-0.99	-	-
$ArticleCount_{t-2}$	-	-0.04	-0.05	-	-
$ArticleCount_{t-3}$	-	0.07	0.16	-	-
$ArticleCount_{t-4}$	-	0.70	0.72	-	-
$ArticleCount_{t-5}$	-	-1.46*	-1.43*	-	-
Neg_t	-	-	-	1.01	1.05
Neg_{t-1}	-	-	-	-0.00	0.04
Neg_{t-2}	-	-	-	0.59	0.61
Neg_{t-3}	-	-	-	-1.35*	-1.29*
Neg_{t-4}	-	-	-	0.59	0.62
Neg_{t-5}	-	-	-	-0.53	-0.49
Vol_t	68.73	15.30	-	18.65	-
Vol_{t-1}	-170.48	-33.71	-	-77.32	-
Vol_{t-2}	-343.87	-431.53	-	-409.19	-
Vol_{t-3}	1362.42	1357.96	-	1422.32	-
Vol_{t-4}	-1043.60	-1061.69	-	-1106.98	-
Vol_{t-5}	132.21	159.40	-	157.618	-
Cal. 1: *Jan*	-2.97	-2.68	-3.13	-3.04	-3.49
Cal. 2: *Rec*	-10.16**	-10.50**	-9.07*	-10.09**	-8.88*
RSS	0.0880	0.0878	0.0879	0.0877	0.0878
R^2	0.0218	0.0234	0.0223	0.0246	0.0235

Coefficient terms shown in basis points (1 BP = 0.01%)
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 4.9: UK Panel Regression- Models 3, 4 and 5 (01/09/2003 - 18/01/2018)

	Model 3	**Model 4a**	**Model 4b**	**Model 5a**	**Model 5b**
			$n = 3555$		
Constant	-32.99	-79.26	0.54	15.15	9.49**
r_{t-1}	-411.03	-417.91	-406.11	-602.72**	-599.91**
r_{t-2}	-538.88	-548.58	-532.12	-491.75	-482.22
r_{t-3}	-481.37	-468.27	-460.56	-485.37	-478.25
r_{t-4}	337.26	314.71	304.62	241.92	231.33
r_{t-5}	-594.85*	-595.78*	-603.05*	-521.53	-529.46
$ArticleCount_t$	-	-0.17	-0.19	-	-
$ArticleCount_{t-1}$	-	0.18	0.14	-	-
$ArticleCount_{t-2}$	-	0.01	-0.02	-	-
$ArticleCount_{t-3}$	-	0.09	-0.10	-	-
$ArticleCount_{t-4}$	-	-0.16	-0.16	-	-
$ArticleCount_{t-5}$	-	-0.29**	0.27*	-	-
Neg_t	-	-	-	-11.78***	-11.78***
Neg_{t-1}	-	-	-	1.41	1.37
Neg_{t-2}	-	-	-	2.31*	2.31*
Neg_{t-3}	-	-	-	1.52	1.67
Neg_{t-4}	-	-	-	-0.74	-0.62
Neg_{t-5}	-	-	-	3.07**	3.15**
Vol_t	-54.06	-103.52	-	73.02	-
Vol_{t-1}	-247.15	-201.80	-	-456.78	-
Vol_{t-2}	807.31	820.24	-	853.57*	-
Vol_{t-3}	-178.14	-167.98	-	-188.07	-
Vol_{t-4}	-320.56	-370.84	-	-252.35	-
Vol_{t-5}	-3.55	32.64	-	-29.94	-
Cal. 1: *Jan*	-2.21	-2.22	-3.40	-2.62	-3.49
Cal. 2: *Rec*	-7.04	-7.15	-6.39	-5.57	-5.83
RSS	0.0836	0.0834	0.0835	0.0814	0.0815
R^2	0.0142	0.0165	0.0146	0.0401	0.0388

Coefficient terms shown in basis points (1 BP = 0.01%)
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 4.10: India Panel Regression- Models 3, 4 and 5 (01/09/2003 - 18/01/2018)

	Model 3	**Model 4a**	**Model 4b**	**Model 5a**	**Model 5b**
			$n = 3446$		
Constant	-42.93	-50.94	3.75**	-20.33	6.81***
r_{t-1}	713.38**	715.82**	718.54**	601.14*	595.21*
r_{t-2}	-559.11*	-554.26*	-553.78*	-544.93*	-548.94*
r_{t-3}	-170.76	-171.52	-193.47	-183.74	-206.56
r_{t-4}	-122.75	-111.28	-100.69	-72.72	-69.88
r_{t-5}	-355.52	-346.08	-336.78	-361.85	-357.27
$ArticleCount_t$	-	0.15	0.02	-	-
$ArticleCount_{t-1}$	-	0.13	0.06	-	-
$ArticleCount_{t-2}$	-	-0.07	-0.07	-	-
$ArticleCount_{t-3}$	-	0.47	0.46	-	-
$ArticleCount_{t-4}$	-	0.40	0.31	-	-
$ArticleCount_{t-5}$	-	-0.74**	-0.86**	-	-
Neg_t	-	-	-	-6.65***	-6.83***
Neg_{t-1}	-	-	-	2.20**	2.12**
Neg_{t-2}	-	-	-	0.13	0.16
Neg_{t-3}	-	-	-	2.64**	2.66**
Neg_{t-4}	-	-	-	-0.38	-0.48
Neg_{t-5}	-	-	-	-0.02	-0.10
Vol_t	73.83	59.42	-	115.10	-
Vol_{t-1}	-87.65	-84.90	-	-153.35	-
Vol_{t-2}	889.34*	886.35*	-	819.76	-
Vol_{t-3}	-1139.35**	-1136.47**	-	-988.92*	-
Vol_{t-4}	-159.90	-123.74	-	-203.33	-
Vol_{t-5}	434.92	412.28	-	417.16	-
Cal. 1: *Jan*	-5.28	-5.22	-5.45	-5.90	-6.00
Cal. 2: *Rec*	-10.83*	-10.95*	-7.53	-10.11	-8.21
RSS	0.1347	0.1345	0.1349	0.1328	0.1331
R^2	0.0144	0.0158	0.0128	0.0281	0.0261

Coefficient terms shown in basis points (1 BP = 0.01%)

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Consistent with the prior findings, it is seen that the proxy variable for economic recession maintains statistical significance in the US market across each model, negatively contributing between 8.9 and 10.5 basis points. As also seen previously, it never holds statistical significance when applied to the UK markets. Contrary to previous findings however, the recession variable does carry statistical significance in the Indian market, negatively contributing between 10.8 and 11 basis points, but this is only exhibited in Models 3 and 4a, and is otherwise washed out.

It is seen that detrended volume has no statistical significance in the US market, whereas the opposite is exhibited in the Indian market, with at least one lag of detrended volume carrying statistical significance in each of the models it is accounted for. Note that its impact on returns is sizable, negatively contributing over 1139 and 1136 basis points on the third lag terms in Models 3 and 4a respectively, and approximately 989 basis points on the third lag in the best fitting model for the Indian market. The UK market exhibits interesting behaviour, with detrended volume carrying statistical significance only when also accounting for the negative sentiment proxy in the best fitting model, negatively contributing 854 basis points on the second lag term.

All three markets show statistical significance of the article count variable on the fifth lag term, however it should be noted that its contribution is consistently small- negatively contributing between 0.3 and 1.5 basis points across all three markets. The negative sentiment proxy also always carries statistical significance, however has a more varied contribution across the three markets. The US market exhibits its statistical significance only on the third lag, negatively contributing 1.35 basis points in the best fitting model (Model 5a). The UK market on the other hand indicates statistical significance of the proxy on the second and fifth lags in the best fitting model, positively contributing 2.31 and 3.07 basis points respectively. Finally, the Indian market exhibits statistical significance of negative sentiment on the first and third lag terms in the best fitting model, positively contributing 2.20 and 2.64 basis points respectively. Interestingly, it should be noted that the best fitting models for the UK and Indian markets additionally suggest that the "same day" term carries statistical significance that- unlike the positive contributions of its lagged counterparts in both markets- negatively contributes a significantly larger amount towards returns (11.78 and 6.65 basis points for the UK and Indian markets respectively).

A further examination of the residual sum squared (RSS) and R^2 performance evaluation metrics for each model across the three markets indicates that the best estimate model for return is Model 5b, which considers lagged return values, detrended volume and the calculated negative sentiment proxy as factors influencing market returns. Thus this study delivers promising results, suggesting that the use of sentiment analysis to estimate the negative sentiment present in formal media delivers potential value when modelling financial return.

4.3.2 Case Study: Recessionary Breakdown

The time period encompassed by this study includes the eighteen-month period from December 2007 to June 2009 [8] that saw one of the biggest downturns in the global economy- the "Great Recession". Hence an opportunity is presented to explore how the results obtained by this study that quantify the impact of negative sentiment on financial return differ when examined over three sub-periods: before, during and after the recession. The aforementioned best estimate model for recession was selected as the model of choice for this analysis, excluding the proxy variables for calendar effects. The proxy variable for economic recession was omitted due to the subdivision of time periods into predefined recessionary and non-recessionary periods. The proxy variable to account for the period following the major Christmas and New Year holidays was omitted due to its consistent statistical insignificance in all previous results. Note that in the following results, a one-month period was included as a buffer around the recessionary period.

Table 4.11 shows the results of the VAR analysis performed on each market for the period before the recession from the beginning of September 2003 until the end of October 2007. It is seen that the coefficient terms for detrended volume has no statistical significance across all three markets. The proxy variable for negative sentiment carries statistical significance only in the Indian market, on the "same day" term, negatively contributing 9.62 basis points towards return. It is also seen that detrended volume is a large contributor to return, however its coefficient values exhibit no statistical significance. Lastly, it is important to observe that a quick examination of the R^2 values across the three markets indicates the model has a considerably better "goodness of fit" when applied to the Indian market than when applied to the US and UK markets.

Table 4.12 shows the results of the VAR analysis performed on each market for the period during the recession from the beginning of November 2007 until the end of July 2009, including a one-month buffer either side of the official start and end dates of the recession provided by the National Bureau of Economic Research [8]. Several interesting observations are made based on the results of this VAR analysis. Firstly, it is noticed that the negative sentiment proxy comes into statistical significance in the US market on the third lag term, negatively contributing 5.61 basis points to return. Similarly, the UK market sees the introduction of statistical significance for the proxy on the "same day" term, negatively contributing a surprisingly large 31.85 basis points. Finally, the Indian market is shown to maintain statistical significance for negative sentiment on the "same day" term, however its negative contribution has increased to 24.65 basis points. Additionally, the first lag term has come into statistical significance, positively contributing 10.52 basis points towards return. Is is important to further note that detrended volume comes into statistical significance in the Indian market during the recession on the "same day", first and third lag terms, with both positive and negative contributions ranging from 2677 basis points to as much as 9751 basis points. Examining the R^2 values across the three markets, it is further noted that the "goodness of fit" of the model for each market is shown to increase significantly when compared to the period prior to the recession.

Table 4.13 shows the results of the VAR analysis performed on each market for the period after the recession from the beginning of August 2009 until mid-January 2018- the end of the time-frame encapsulated by the study. Immediately it is noticed that the US market loses the statistical significance it had carried on negative sentiment during the recessionary period. We see that statistical significance is still maintained in the UK market, and although it's contribution is lessened, the significance extends from what was previously just of the "same day" term (whose negative contribution decreased to 19.09 basis points) to also include the first, third and fifth day lags, positively contributing 4.37, 3.45 and 6.01 basis points respectively. This is a surprising outcome, however it is worth noting that other significant economic events that took place within the time period of interest not considered by this study may play a potential role here- a primary suspect candidate being the economic fallout resulting from the UK vote to leave the European Union in 2016. The negative sentiment proxy is shown to maintain its statistical significance in the Indian market, with a negative contribution of 2.11 and 1.65 basis points on the "same day" and fourth lag terms respectively, and a positive contribution of 1.54 basis points towards return on the third lag term. Additionally, note that the fourth lag term for detrended volume now carries statistical significance, negatively contributing 867.59 basis points to return. Finally, an examination of the R^2 terms for the model across the three markets indicates an overall reduction in "goodness of fit" compared to the period during the recession.

Table 4.11: VAR Analysis Before Recession (01/09/2003 - 31/10/2007)

	USA	**UK**	**India**
	$n = 985$	$n = 978$	$n = 958$
Constant	-28.85	296.90	141.55*
r_{t-1}	-709.75**	-1230.70**	354.60
r_{t-2}	-608.67*	-26.16	-1099.29**
r_{t-3}	365.32	-217.09	292.91
r_{t-4}	-365.33	-112.09	850.71*
r_{t-5}	-46.26	22.47	-173.40
Vol_t	1078.55	-321.41	-339.42
Vol_{t-1}	-2005.54	497.09	274.89
Vol_{t-2}	815.11	399.35	412.57
Vol_{t-3}	89.74	-1103.87	-614.74
Vol_{t-4}	854.63	1152.31	883.16
Vol_{t-5}	-829.40	-655.48	-647.12
Neg_t	-0.02	-1.75	-9.62***
Neg_{t-1}	0.38	0.05	0.66
Neg_{t-2}	0.53	0.17	0.36
Neg_{t-3}	1.47	-1.10	2.56
Neg_{t-4}	-0.24	0.96	-0.31
Neg_{t-5}	-0.24	1.52	-0.78
RSS	0.0094	0.0105	0.0374
R^2	0.0210	0.0236	0.0474

Coefficient terms shown in basis points (1 BP = 0.01%)
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 4.12: VAR Analysis During Recession (01/11/2007 - 31/07/2009)

	USA	**UK**	**India**
	$n = 435$	$n = 437$	$n = 416$
Constant	-518.76	-253.66	-282.66*
r_{t-1}	-1457.76**	-1040.11*	619.45
r_{t-2}	-1317.94	-770.93	-314.69
r_{t-3}	695.67	-984.45	-570.88
r_{t-4}	-346.82	984.70	-690.78
r_{t-5}	-306.16	-890.34	-599.68
Vol_t	2188.79	-609.16	-2677.40*
Vol_{t-1}	-2933.84	-1613.35	4325.12*
Vol_{t-2}	-2233.22	3883.62	3462.94
Vol_{t-3}	6237.39	394.33	-9750.99**
Vol_{t-4}	-3874.25	-1923.78	3576.55
Vol_{t-5}	668.57	-99.49	1127.25
Neg_t	2.71	-31.85***	-24.65***
Neg_{t-1}	1.36	5.49	10.52*
Neg_{t-2}	-01.33	7.60	1.41
Neg_{t-3}	-5.61*	6.11	7.06
Neg_{t-4}	1.75	-7.82	3.89
Neg_{t-5}	-2.40	0.38	-0.13
RSS	0.0418	0.0312	0.0514
R^2	0.0697	0.1305	0.1138

Coefficient terms shown in basis points (1 BP = 0.01%)
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 4.13: VAR Analysis After Recession (01/08/2009 - 18/01/2018)

	USA	**UK**	**India**
	$n = 2125$	$n = 2130$	$n = 2062$
Constant	-20.14	-93.34	-4.29
r_{t-1}	-463.13	-47.14	577.79**
r_{t-2}	108.07	-334.12	-38.24
r_{t-3}	-592.00*	12.12	-276.29
r_{t-4}	-164.36	-769.62***	-164.70
r_{t-5}	-682.73**	-189.11	-102.64
Vol_t	-363.49	288.20	328.83
Vol_{t-1}	431.15	-352.93	-604.43
Vol_{t-2}	-44.90	314.59	721.74
Vol_{t-3}	571.43	-249.10	-44.48
Vol_{t-4}	-834.85	-167.23	-867.59*
Vol_{t-5}	242.96	177.94	467.86
Neg_t	0.66	-19.09***	-2.11**
Neg_{t-1}	-0.34	4.37**	1.13
Neg_{t-2}	0.88	2.38	0.07
Neg_{t-3}	-0.78	3.45*	1.54*
Neg_{t-4}	0.17	-0.64	-1.65*
Neg_{t-5}	-0.05	6.01***	-0.10
RSS	0.0343	0.0355	0.0381
R^2	0.0151	0.0601	0.0162

Coefficient terms shown in basis points (1 BP = 0.01%)

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

5 Conclusion

This chapter serves as a summary of the work outlined over the course of the study, and provides an overview of the key findings across the US, UK and Indian markets during periods before, during and after economic recession. Finally, the limitations of the methods employed by this study are outlined, and suggestions for future work building on this study are proposed.

5.1 Summary and Findings

This study focused on examining the impact of negative sentiment present in formal media- across print and web-based publications- on financial return in the United States of America, United Kingdom and India. This was systematically accomplished through three distinct phases of work: data acquisition and preprocessing, text analysis and statistical modelling. The first phase involved the gathering of financial time series data and media content across the three markets over a period of close to fifteen years. In the second phase, negative sentiment that was present in the retrieved media content was identified and extracted through a dictionary-based approach to sentiment analysis, resulting in the computation of a negative sentiment proxy time series for each market. In the final phase, the financial time series including return, detrended volume and calendar proxy variables was jointly analyzed with the negative sentiment proxy time series using a process known as vector autoregression, in order to model financial return for each market. The VAR models generated facilitated the computation of return based on the other variables accounted for by the system, and further estimated the linear dependence of return on each of these variables. Thus it was possible to investigate the impact of negative sentiment on financial return for each of the three markets in question.

The results obtained from the study indicate that that sentiment has a measurable impact on return that is episodic and time-varying in nature, carrying statistical significance across each of the three markets. Furthermore, it is observed that the statistical models that account for negative sentiment exhibit a better fit for the data than the models that omit the proxy- thus suggesting that the consideration of sentiment as an influencing factor on return is justified and adds value to the analysis of financial return overall.

The extended study over a breakdown of the periods before, during and after the economic recession yields further interesting insights. The US market shows that the negative sentiment proxy is only statistically significant during the period of recession, and is otherwise not. This suggests that the sentiment tends to have a significant impact on return only during a period of uncertainty and volatility, and upon the end of such a period, this is washed out as the market settles. The UK market shows no statistical significance of the negative sentiment proxy prior to the recession. This changes upon the onset of the recession, and is surprisingly maintained (albeit with reduced contribution) through into the period afterward. As a well established market in a developed economy, similar to that present in the United States, the continued significance of negative sentiment in the period after the recession contrasts the findings from the US market. However as previously discussed, it is important to consider that perhaps economic events not accounted for by this study- such as the fallout resulting from the UK vote to leave the European Union in 2016- contributed to this result. Lastly, the Indian market consistently indicates statistical significance of the negative sentiment proxy- even in non-recessionary periods- although its contribution is shown to have heightened considerably during the recessionary period. This suggests a higher and continuous degree of uncertainty present in the more recently established and emerging market, than present in its US and to an extent, UK counterparts. A crucial observation across these sub-periods is that the statistical models have a considerably better "goodness of fit" when applied to the period during the recession. This suggests that perhaps sentiment is predominantly influential during periods of high market uncertainty and volatility.

5.2 Future Work

The contributions of this study include a developed system that quantifies the sentiment present within a corpus of texts, and investigates its influence on financial markets through multivariate time series analysis. The system successfully shows that sentiment has a measurable impact on return that is episodic and time-varying in nature. While proven techniques for sentiment analysis and multivariate time series analysis are employed by this system, there are limitations of this study that may be addressed by future work. The developed system examines return and sentiment at a daily frequency, considering the daily adjusted closing index values, and grouping articles by day. Considering the trend of high frequency trading in today's markets, an analysis performed at a higher frequency would facilitate a more granular study of the effects of sentiment, and a more in-depth study of the volatility of the data.

The study provides justification for the use of a dictionary-based approach to sentiment analysis in conjunction with a "bag-of-words" text representation model, however the approach has not been evaluated over the course of the work. Use of different lexicons for example would facilitate a comparison across these lexicons and an analysis of the effectiveness of the constructed glossary as previously detailed. The approach is also somewhat limited in nature, due to its requirement for all affect terms to be included in the predefined lexicon, thus demanding a knowledge of representative affect terms for the domain in question. Machine learning approaches on the other hand are equipped with the ability to learn these affect terms through training, testing and validation. Their use for a study of the same markets with similar financial data and media content, over the same period of time, would provide comparable results which could be used to evaluate the comprehensiveness of this dictionary-based approach to sentiment analysis.

Bibliography

[1] Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan, and Maitreyee Dutta. Sentiment analysis for indian stock market prediction using sensex and nifty. *Procedia Computer Science*, 70:85 – 91, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.10.043. URL `http://www.sciencedirect.com/science/article/pii/S187705091503207X`. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.

[2] Henry G. Green and Michael A. Pearson. Artificial intelligence in financial markets. volume 2, pages 839 – 844, Perth, Aust, 1995. Financial market trading;.

[3] Khurshid Ahmad. Edderkoppspinn eller nettverk: New media and the use of polar words in emotive contexts. *Synaps*, 21:20 – 36, 2008. URL `https://www.scss.tcd.ie/khurshid.ahmad/Research/Sentiments/2008_Sentiment_Religion_Synapse.pdf`.

[4] Khurshid Ahmad, JingGuang Han, Elaine Hutson, Colm Kearney, and Sha Liu. Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance*, 37:152 – 172, 2016. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2015.12.014. URL `http://www.sciencedirect.com/science/article/pii/S0929119915001637`.

[5] Stephen Kelly. *News, Sentiment, and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets*. PhD thesis, University of Dublin, Dublin, October 2016. URL `http://www.tara.tcd.ie/handle/2262/79727`.

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

[7] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[8] NBER. Us business cycles expansion and contraction, September 2010. URL `http://www.nber.org/cycles/sept2010.html`.

[9] Eugene F Fama. The behavior of stock-market prices. *The journal of Business*, 38(1): 34–105, 1965.

[10] Jonathan Clarke, Tomas Jandik, and Gershon Mandelker. The efficient markets hypothesis. *Expert financial planning: Advice from industry leaders*, pages 126–141, 2001.

[11] Jyoti Kumari and Jitendra Mahakud. Does investor sentiment predict the asset volatility? evidence from emerging stock market india. *Journal of Behavioral and Experimental Finance*, 8:25 – 39, 2015. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2015.10.001. URL `http://www.sciencedirect.com/science/article/pii/S2214635015000593`.

[12] Andrei Shleifer and Lawrence H Summers. The noise trader approach to finance. *Journal of Economic perspectives*, 4(2):19–33, 1990.

[13] Ian Harvey. Noise traders. URL `https://www.investopedia.com/university/introduction-stock-trader-types/noise-traders.asp`.

[14] J Bradford De Long, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738, 1990.

[15] Malcolm Baker, Joshua Coval, and Jeremy C Stein. Corporate financing decisions when investors take the path of least resistance. *Journal of Financial Economics*, 84(2): 266–298, 2007.

[16] Albert Phung. Behavioral finance. URL `https://www.investopedia.com/university/behavioral_finance/`.

[17] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

[18] Zeyan Zhao and Khurshid Ahmad. Qualitative and quantitative sentiment proxies: interaction between markets. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 466–474. Springer, 2015.

[19] Zeyan Zhao and Khurshid Ahmad. A computational account of investor behaviour in chinese and us market. *Int. J. Econ. Behav. Organ*, 3(6):78–84, 2015.

[20] Kenneth L Fisher and Meir Statman. Investor sentiment and stock returns. *Financial Analysts Journal*, 56(2):16–23, 2000.

[21] Jussi Nikkinen and Sami Vähämaa. Terrorism and stock market sentiment. *Financial Review*, 45(2):263–275, 2010.

[22] Lee Gillam, Khurshid Ahmad, Saif Ahmad, Matthew Casey, David Cheng, Tugba Taskaya, PCF de Oliveira, and Pensiri Manomaisupat. Economic news and stock market correlation: A study of the uk market. In *Conference on Terminology and Knowledge Engineering*, 2002.

[23] Saumya Ranjan Dash and Debasish Maitra. Does sentiment matter for stock returns? evidence from indian stock market using wavelet approach. *Finance Research Letters*, 2017. ISSN 1544-6123. doi: https://doi.org/10.1016/j.frl.2017.11.008. URL `http://www.sciencedirect.com/science/article/pii/S1544612317305111`.

[24] Debasish Maitra and Saumya Ranjan Dash. Sentiment and stock market volatility revisited: A time–frequency domain approach. *Journal of Behavioral and Experimental Finance*, 15:74 – 91, 2017. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2017.07.009. URL `http://www.sciencedirect.com/science/article/pii/S2214635017300515`.

[25] Ronald Hochreiter. Computing trading strategies based on financial sentiment data using evolutionary optimization. In *Mendel 2015*, pages 181–191. Springer, 2015.

[26] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[27] Susan Dumais et al. Using svms for text categorization. *IEEE Intelligent Systems*, 13 (4):21–23, 1998.

[28] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[29] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[30] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.

[31] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[32] Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558, 2009.

[33] Vineet Yadav, Harsha Elchuri, et al. Serendio: Simple and practical lexicon based approach to sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 543–548, 2013.

[34] Cataldo Musto, Giovanni Semeraro, and Marco Polignano. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. *Information Filtering and Retrieval*, 59, 2014.

[35] M. Pant, K. Deep, A. Nagar, and J.C. Bansal. *Proceedings of the Third International Conference on Soft Computing for Problem Solving: SocProS 2013*. Number v. 2 in Advances in Intelligent Systems and Computing. Springer India, 2014. ISBN 9788132217688. URL `https://books.google.ie/books?id=AJu6BQAAQBAJ`.

[36] Antonio Moreno-Ortiz and Javier Fernández-Cruz. Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. *Procedia-Social and Behavioral Sciences*, 198:330–338, 2015.

[37] Basant Agarwal and Namita Mittal. Semantic orientation-based approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis*, pages 77–88. Springer, 2016.

[38] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[39] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[40] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26, 2007.

[41] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[42] Khurshid Ahmad and Andrea Zemánková. Methods and system for calculating affect scores in one or more documents, September 18 2014. US Patent App. 14/214,080.

[43] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[44] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. 2009.

[45] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307, 2011.

[46] Paul F Barba, Michael W Marshall, and Carl J Lambrecht. Methods for analyzing text, May 10 2016. US Patent 9,336,192.

[47] Gautam Mitra and Leela Mitra. *The handbook of news analytics in finance*, volume 596. John Wiley & Sons, 2011.

[48] David Leinweber and Jacob Sisk. Event driven trading and the'new news'. 2011.

[49] Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912017.

[50] Christopher A. Sims. Statistical modeling of monetary policy and its effects. *American Economic Review*, 102(4):1187–1205, 2012.

[51] Khurshid Ahmad. Mood changes and their impact on the value attached to goods, services and people: the role of verbal and visual sentiment and the 'market place'. 2017.

[52] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[53] Chung Baek. Stock prices, dividends, earnings, and investor sentiment. *Review of Quantitative Finance and Accounting*, 47(4):1043 – 1061, 2016. ISSN 0924865X. URL http://search.ebscohost.com.elib.tcd.ie/login.aspx?direct=true&db=eoh&AN=1610444&site=ehost-live.

[54] Sharat Chandran. How the sensex is calculated, February 2008. URL http://www.rediff.com/money/2008/feb/21bspec.htm.

[55] Bloomberg. Sensex, March 2018. URL https://www.bloomberg.com/quote/SENSEX:IND.

[56] Sanam Mirchandani. What is free float market capitalisation, July 2017. URL https://economictimes.indiatimes.com/markets/stocks/news/free-float-market-capitalisation-determines-index-weightage/articleshow/58178327.cms.

[57] John Y Campbell, Sanford J Grossman, and Jiang Wang. Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, 108(4):905–939, 1993.

[58] Economist. Economics a-z terms, April 2018. URL `https://www.economist.com/economics-a-to-z`.

[59] Bishop Matthew. *Essential economics*. Economist in association with Profile Books, 2004.

[60] business.gov.au. Key financial terms, April 2018. URL `https://www.business.gov.au/info/run/finance-and-accounting/finance/key-financial-terms`.

[61] IIFL. About iifl, April 2018. URL `https://www.iifl.com/about-us`.

[62] indiainfoline.com. Stock market glossary, April 2018. URL `https://www.indiainfoline.com/article/research-articles-ideas/stock-market-glossary-113111500414_1.html`.

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

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

[65] Stephen J. Taylor. *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press, stu - student edition edition, 2005. ISBN 9780691134796. URL `http://www.jstor.org/stable/j.ctt7t66m`.

[66] Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix, 1986.

[67] Whitney K Newey and Kenneth D West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.

[68] Stephen J Brown and Jerold B Warner. Using daily stock returns: The case of event studies. *Journal of financial economics*, 14(1):3–31, 1985.

[69] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Heiko Ebens. The distribution of realized stock return volatility. *Journal of financial economics*, 61(1):43–76, 2001.