News, Sentiment, and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets.

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Stephen Kelly
October 2016

School of Computer Science and Statistics
Trinity College Dublin

To my Parents and sister Rachel for their endless love and support.

Declaration

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

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

Signed,

Stephen Kelly

October 12, 2016.

Abstract

With the advent of the internet and digitisation of news and books, the volume of unstructured text has increased dramatically in recent years. This deluge of information is set to grow and come from new and unconventional sources. New and innovative techniques and tools for manipulating this data and making sense of it will become essential. The work in this thesis consists of a system that analyses the content of news, extracting a sentiment time series variable, and uses this variable in a time series modelling component to determine any inter-relationships between changes in the price of financial assets. Each component of the system attempts to remove human subjectivity from the modelling process to allow the system to compute a sentiment variable and investigate its statistical significance and explanatory power by employing several time series models. The system includes a number of processes and components to achieve this goal such as data harvesting, processing, text analysis, time series modelling, hypothesis testing, and visualisation. The work described in this thesis includes contributions to the area of text and content analysis, information retrieval, time series analysis, statistical and econometric modelling.

A number of methods studied in the literature have incorporated text data with financial analysis and prediction models. A review of some of the main studies and systems that have combined methods from each discipline is presented in Chapter 2. Chapter 3 describes the system developed in this thesis and its capabilities. The system is evaluated for different data inputs in Chapter 4, where the influence of different news types and sources is investigated for equity and commodity markets. The final chapter concludes the thesis by summarising the contributions and outlining future work.

This thesis represents a combination of work that contributes to the areas of content analysis and financial and statistical modelling. The main contribution of this thesis is in the implementation and evaluation of a system that incorporates methods from text analysis and time series modelling. The novelty of the system lies in the ability to compute a time series from text that acts as a proxy for sentiment that can be aggregated with financial time series data in a statistical model to estimate the impact of news on the financial asset. An evaluation of this system is presented where the explanatory power of the sentiment variable for financial returns is investigated. It is shown that the news source and text type play an important role when computing a proxy for sentiment that has statistically significant explanatory power for financial returns. The time varying influence of sentiment on financial returns is noted with a relationship made between economic business cycles and volatility. The system is evaluated using data from two financial markets, the equity and commodities markets, and in both instances it is found that a proxy for sentiment extracted from news has a statistically significant influence on financial returns.

Acknowledgements

"All truths are easy to understand once they are discovered; the point is to discover them"

- Galileo Galilei

I would like to thank a number of people who have helped make this dissertation possible. I would like to express my gratitude and sincere thanks to Professor Khurshid Ahmad for his supervision. Without his continued support, encouragement, and expertise this thesis would not have been possible. I would also like to thank the hospitality of Trinity College Dublin, the School of Computer Science and Statistics, and the Faireachain project (IP/2009/0595) for scholarship support.

Thanks are due to all those, too many to mention, who have passed through 116 and that I have worked with during this long period of immersive study. Thanks also to Tomas Dunne for his words of encouragement throughout. Gratitude must be given to Ian Kelly, who I'm glad was also on the same path and whose topics of discussion on long walks were edifying and greatly beneficial.

Most of all I am eternally greatly to my family, my Mam, Dad, and sister Rachel, who without their unconditional love and support I would most certainly not be here today. I cannot express my gratitude enough and for reminding me this was a good idea. I will pay you all back with interest.

Lastly, I would like to thank Immy who I love very much for her absolute patience, keeping me motivated throughout and always making me smile.

Contents

List of Acronyms

ACF Autocorrelation Function

ADF Augmented Dickey-Fuller test

AOTM Abreast of the Market

API Application Programming Interface

AR Autoregression

ARCH Autoregressive Conditional Heteroskedasticity

DJIA Dow Jones Industrial Average

FT Financial Times

GARCH Generalised Autoregressive Conditional Heteroskedasticity

OLS Ordinary Least Squares

PACF Partial Autocorrelation Function

VAR Vector Autoregression

VSM Vector Space Model

WSJ Wall Street Journal

WTI West Texas Intermediate

1 Introduction

1.1 Motivation and Background

It is natural to assume that decisions and opinions can be influenced by news found in printed, digitised and televised media, as well as by word of mouth. How news and the underlying information influences behaviour has become a topic of interest in social, economic, and computational studies. The flow of information and uncertainty in markets influences price discovery, price changes, volatility, the behaviour of market participants, and market stability. A complex relationship exists between news and markets where the arrival of news may change investors' beliefs, update their knowledge of the market, and ultimately influence sentiment.

With advancements in the reach and speed of information, access to quantitative and qualitative information has become faster and inexpensive. Traditional news has evolved and become digitised and has a greater range of dissemination being published online through websites, blogs, and social media. With this increase in the channels of communication, investors in financial markets can now access information faster than ever before. News and reports are available in text format online and provided through vendors such as Reuters and Bloomberg. These news stories represent a wealth of information that is often not fully reflected in financial market data. Increased interest has been seen in the finance community to employ methods of text and content analysis to automatically analyse this information in the hope of adding explanatory information to traditional financial models. Combining methods from these two disciplines, that of text analysis and econometric analysis, has only in recent times garnered much attention [8] [42] [96], with systems being developed to automate this process and fully integrate the methods and data

processing tasks [23] [43] [88]. Creating a system to handle these vast amounts of news text and compute a model that can utilise this information poses an interesting challenge.

News articles are an important part of market information and are widely read by investors and market participants. The behaviour, opinions, and mood of some investors, often referred to as investor sentiment in the domain of behavioural finance, is said to be linked to the timing and publication of news. In recent times, studies in the domain of finance have turned to using text analysis to examine the content of news and to create a proxy for investor sentiment [42] [65] [66] [96] [97]. Text analysis has been used in a number of studies to extract information from a number of different text documents such as annual reports, company filings (10-Ks, 10-Qs), earnings reports, analyst reports, internet message boards, and social media. A theoretical model of investor sentiment may differ from what textual sentiment may represent. Investor sentiment might encompass subjective beliefs and behaviour. Text based sentiment may only represent how investors react to news and the text content. Only tentative assumptions can be made about the link between investor behaviour and sentiment summarised from news and text. Despite this, the quantification of this otherwise qualitative information gives a new and consistent complementary information source. The inclusion of textual sentiment in predictive models of returns and price gives an alternative approach to incorporating new information into traditional quantitative models.

Text and content analysis methods have been used to measure the tone of a text and the output of such methods and systems have been used in inferential and predictive models particularly in the domain of finance. Studies in the literature have used methods of text analysis, machine learning, econometrics, and statistical modelling in isolation. Fewer studies have created systems or approaches that have attempted to combine these methods together. Of the studies conducted and systems created to do both text analysis and statistical modelling, few if any have created a framework that can be generalised to allow the analysis of different text sources, types, financial, and time series data.

One of the present challenges lies in combining econometric and statistical methods with that of text and content analysis. A proxy for information derived from news whether evidence of investor sentiment or not, may highlight anomalies such as potential mispricing of an asset, agreement or disagreement in a market, or just a response of the market to news being released. A proxy for news sentiment can be beneficial and improve the explanatory power of a statistical model of a financial asset. The question then becomes how to analyse the content of news to produce a sentiment variable that has explanatory information and how this variable can be incorporated into a statistical model to determine the impact of news on the price, returns, or level of a financial asset.

The main objective of this thesis is the development and evaluation of a system that can perform these tasks, allowing for different data inputs and models to be used interchangeably and combined together. The system aims to combine methods from the area of text analysis and time series modelling to process and analyse the sentiment in different news text to determine

a relationship with movements in assets traded on financial markets. To fulfil this objective, functionality to process and collect a large volume of mostly unstructured data is required. These functions include harvesting of data from online sources for both financial time series data and text data in structured and unstructured formats. The text analysis method must be capable of structuring and ordering text in order to analyse the content and create a time series representation of the information extracted and to aggregate this with financial time series data. The aggregated time series data needs to be integrated into a statistical model and the inter-relationships between the variables estimated. This framework allows a system to compute a time series of sentiment extracted from a corpus of news and to use this variable in a statistical time series model to estimate changes in the price of a financial instrument.

1.2 Key Conclusions

The focus of this thesis is on creating a computational solution and implementation for computing a proxy for information contained in news content and to evaluate the influence of this proxy on changes in financial markets. Typically, methods of text and sentiment analysis have been used in financial literature to generate this proxy from news. The studies and systems developed to perform this task start by choosing a financial market to analyse, with stocks and indices being the most widely studied. A method or system of summarising the sentiment in a chosen text collection is employed. The choice of data is often subjective and left to the justification of the investigator. This brings the discussion to the development of a system that can allow different data inputs and different models to be used interchangeably. A discussion of the merits of some of these approaches in the literature is given here, where the text analysis and data modelling approaches are reviewed (Chapter 2). This literature review informs and motivates the development of a system and the necessary functionality needed to carry out the processing and analysis tasks.

The system developed in this thesis encompasses an array of tasks including data harvesting and scraping, data processing and aggregation, content analysis, time series modelling, model diagnostics and hypothesis testing, and visualisation. The content analysis method and framework chosen for the implementation allows any text corpus and financial time series data to be collected and imported into the system to estimate the impact of news on financial returns. The system consists of two main components to perform this task, a content analysis component and a statistical modelling component. The first component allows text to be collected and structured into a time ordered corpus, the content of text can be analysed by importing a structured dictionary of terms or a word list of domain terms, which then produces a time series of sentiment. This is then passed to the statistical modelling component, which aligns and aggregates the sentiment with financial data that are retrieved from online data sources. A number of transformations are performed on the time series data which are then passed to a modelling component that performs regression based analysis and hypothesis testing among

other computations. The modelling component evaluates the sentiment variable produced by estimating the inter-relationship between the sentiment variable and financial returns using vector autoregression, a rolling regression model to evaluate the impact of sentiment in time, and a hypothesis test to determine the statistical significance of the sentiment variable and if it contains explanatory information for asset returns. The results can be output and saved, while a web based graphical user interface allows the time series data to be visualised. The system in its implementation facilitates a systematic way of extracting a sentiment variable from different news sources and incorporating this variable into a robust statistical estimation to help explain the movements of financial assets. By using the system to analyse the content of news, a new explanatory variable can be computed and used in traditional financial models (Chapter 3).

The system and underlying methods of text and statistical analysis are evaluated in a number of ways. Two case studies are presented in this thesis that evaluate the system's functionality, modelling capabilities, and output by using data from two different financial markets and using news text from different online sources. The system is evaluated first for the stock market using the Dow Jones Industrial Average and news from authoritative publications. The second case study evaluates the system using data from the commodities market with the oil benchmark West Texas Intermediate Crude oil and news from authoritative publications and blogs. In each investigation the explanatory information in the sentiment variable computed by the system is evaluated. The influence that different text types, and various sources, has on the sentiment variable computed is also evaluated for each financial market in the two case studies. Financial markets and assets are known to be dynamic where the parameters of price distributions (mean, variance) and trends are known to change in time. The same could be true for the influence of the sentiment variable estimated from news by the system. This leads to an investigation into the dynamic time varying impact of the sentiment variable in time by using the system's modelling component and a rolling window method implemented. Finally, a link is made between volatility in financial returns, the sentiment variable and news volume (Chapter 4).

The method used to generate the proxy for sentiment is evaluated for two financial asset classes in two separate case studies. The first case study is carried out on equity markets (using the Dow Jones Industrial Average) and the second for commodity markets (using West Texas Intermediate crude oil). The methods of content analysis are used to extract a sentiment time series, which is then combined with statistical methods drawn from econometrics and statistics to assess the statistical significance.

Initial findings show that negative sentiment extracted from a popular column in the Wall Street Journal, the *Abreast of the Market*, has explanatory power for returns of the Dow Jones Industrial Average index, a result that verified previous findings in the literature. An increase in negative sentiment is found to account for a 4.7 basis point decrease in returns for the Dow Jones Industrial Average. The results suggest that sentiment extracted from news plays a role in explaining changes in the Dow Jones returns. To assess if the source of text plays an important role on creating a proxy for sentiment from news, general business news from the Wall Street

Journal and a collection of text consisting of the *Lex* column from the Financial Times are used to create a proxy for sentiment. A small statistically significant impact is reported for general business news, while the *Lex* column, a similar agenda setting column to *Abreast of the Market*, has up to an 8.5 basis point negative impact on DJIA for negative sentiment extracted from the column. This result is evaluated for a different asset class using the commodity benchmark West Texas Intermediate crude oil and a corpus of oil related news from the Financial Times. The negative sentiment variable negatively impacts the West Texas Intermediate returns, with up to -8.5 basis points being explained by the leading significant coefficient for futures contracts and -9.4 for spot returns. Having determined the average effect of sentiment on different asset classes, the influence of sentiment over time is examined. It is found that the influence of sentiment on the asset classes also changes over time. Using a rolling regression model, it is found that 21% of time series models are significant for sentiment as an explanatory variable for DJIA returns. For commodities, sentiment is a good explanatory variable 13% of the time. These results are compared with the NBER business cycles of recession and expansionary periods finding sentiment has greater impact on WTI futures during recessionary periods (up to -17.3 basis points) than expansionary periods (-6.3 basis points). The evaluation of the content analysis method shows the influence that different text types have on computing the sentiment variable. Three text collections are used, the first consisting of crude oil related news from the Financial Times, the second a dedicated Oil and Gas news section in the Financial Times, and lastly an authoritative industry blog *The Oildrum*. The difference in explanatory information of the negative sentiment proxy extracted from each is evaluated. The general language dictionary, a resource necessary for content analysis, and its use with domain specific text (oil industry news) is also examined. The sentiment proxy is extracted from the domain text while considering potential misinterpretations made by the general language dictionary by using a glossary of domain words and phrases. A difference is seen for the crude oil related text collections from the Financial Times, while the explanatory power of the sentiment proxy from the Oildrum blogs is increased (4.83 basis point increase) for WTI returns.

The contributions of this thesis lay in the method and system implementation of these methods that are used to construct the sentiment proxy and evaluate its effectiveness as an explanatory variable in a statistical model. Previous authors have studied equity markets and text-based sentiment; fewer studies have looked outside this market when evaluating sentiment. This thesis evaluates the use of content analysis in equity markets in the first of the two case studies. The second case generalises the approach and applies it to a different market, that of commodities. Fewer studies have focused on commodity markets and sentiment from oil news. Results presented in this study verify and evaluate the method and system and find a similar significant impact of news sentiment on the returns of the crude oil benchmark. The use of a computational system allows the method described in this thesis to be generalised allowing any text input and financial data to be input, assessing the impact of news on financial returns.

1.3 Publications

The groundwork for this thesis was first published in Kelly and Ahmad [55] where a proxy for a weekly announcement was created. The change in price around the weekly announcement of crude oil inventory was sampled around the period of the announcement. This change was used as a proxy for the announcement and incorporated into a basic regression model as a perturbation, as was textual sentiment. The model demonstrated an increase in predictive power over using a typical time series dummy variable for the announcement time:

Kelly, Stephen, and Khurshid Ahmad. "Sentiment proxies: computing market volatility." *Intelligent Data Engineering and Automated Learning-IDEAL 2012. Springer Berlin Heidelberg, 2012. 771-778.*

The method of constructing a time series for affect from text was also employed in:

Kelly, Stephen, and Khurshid Ahmad. "Determining levels of urgency and anxiety during a natural disaster: Noise, affect, and news in social media." *DIMPLE: DIsaster Management and Principled Large-scale information Extraction Workshop Programme.*

Methods of econometrics, text analysis, and proxy determination have also been included in the following preprint article:

Murphy, Tommie, Stephen Kelly, and Khurshid Ahmad. "Innovations in the Crude Oil Market: Sentiment, Exploration and Production Methods." *SSRN: Exploration and Production Methods (April 14, 2015)*

Event detection methods for time series and content analysis techniques were employed in:

Kelly, Stephen, and Khurshid Ahmad. "Propagating disaster warnings on social and digital media" *Intelligent Data Engineering and Automated Learning-IDEAL 2015. Springer International Publishing, 2015. 475-484*

A comparison and evaluation of financial and text based proxies was also presented in:

Kelly, Stephen, and Khurshid Ahmad. "The Impact of News Media and Affect in Financial Markets" *Intelligent Data Engineering and Automated Learning-IDEAL 2015. Springer International Publishing, 2015. 535-540*

1.4 Thesis Structure

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

methods of text analysis and financial analysis and prediction are then discussed (Chapter 2).

A detailed description of the system and its implementation is described in Chapter 3. The main components of the system, the text analysis component and the statistical modelling component and the necessary data harvesting and preprocessing functionality are described.

The system is then evaluated by conducting two case studies with text and time series data for two financial markets (Chapter 4). By employing statistical methods such as vector autoregression and hypothesis testing, the explanatory information of the sentiment variable computed by the system from the text corpus for the different financial assets is assessed. How this sentiment variable impacts financial returns is investigated in two financial markets, equities and commodity markets. The influence that different text types and text corpora have on the content analysis method are also evaluated by looking at the role different text sources and news article types play on the computation of the sentiment time series variable. Lastly, an investigation into the changing influence of sentiment over time and in different market conditions is examined for each financial market in the case studies.

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

2

Background and Related Work

2.1 Introduction

The volume of textual data that has grown and become available through the internet in recent times has resulted in many industries being affected. With recent technological advances enabling better storage and retrieval of this information, industries have moved to take advantage of the increased availability of information in order to help inform decision-making and reveal insights. This is particularly true for the domain of finance where distilling the information in digitised text for use in quantitative models has seen increased interest in recent times. The following chapter explores the literature and methods of text analysis used to extract information from text and methods of statistical and time series analysis used to model financial data. The studies that are discussed are from areas such as text and content analysis, econometrics and time series analysis. Some studies covered also include systems and frameworks that draw on methods from both computer science and finance disciplines to build models and assess the impact of information on different classes of assets. One of the objectives of this literature review is to highlight and describe systems and approaches that combine methods of text analysis with statistical models using financial market data.

The literature reviewed in this chapter includes studies that were published on the subject of news and sentiment analysis in finance between the period of 1998 to 2015. More than half of the studies reviewed use sentiment and asset values on a daily time frequency, some vary between using intraday (three studies) and quarterly data (four studies). Sentiment is typically extracted from newspapers (Wall Street Journal) and from news wire sources (Dow Jones News

Wire, Reuters). The focus of these studies is on computing sentiment by identifying single word tokens and matching these tokens with a pre-categorised set of terms in a digital dictionary. This information extraction method is referred to as the bag-of-words model. The manner in which sentiment is aggregated with asset values is often through a regression based analysis (11 studies), which ranges from ordinary least squares (5 studies), Vector autoregression (1) to Fama Macbeth regression (2). A number of authors use machine learning techniques (6), where a training data set of sentiment and asset values is used to establish a causal relationship between the two variables, an example of which is identifying word clusters that are correlated with a rise or fall in asset values.

This chapter will begin by first describing methods of text processing and analysis and their domain of application (Section 2.2). A more detailed discussion of content analysis is then given, as this is the chosen method of text analysis in this work (Section 2.2.2). The news type, sources, and topics of conversation in text all have a considerable influence on the information extracted from a corpus or collection of text. A section discussing corpus construction and lexical resources is given special attention. Models for performing time series and financial analysis are then described including linear regression based models, which can be used to estimate the relationship between different variables (Section 2.3). The last section of this chapter (Section 2.4) discusses studies that have combined methods and from both of these disciplines to create systems and frameworks to assess the influence of text based information on financial markets.

2.2 Text Analysis and Corpus creation

This section will discuss methods of text analysis with an emphasis on content analysis as this is the method that is used in this thesis. Text analysis systems are discussed briefly and their use in the literature. The construction of a corpus from a collection text and the implications that the choice of text can have for different studies is also reviewed.

Observing the frequency and repetition of terms in text is fundamental to content analysis. The conjecture made is that by counting the frequency of affect or sentiment terms in text, it may be possible to summarise the tone of a text and represent the sentiment present in the text. When examining text, the repetition of terms must be relied upon to provide consensus. Whether these terms represent any meaning or are well understood is an agreement that is intuitively understood by native speakers of a language [84].

> *Assessments by native speakers of relative acceptability largely correlate with their assessments of relative frequency. (Quirk et al., 1985) [84]*

Acceptability in this context is a wider concept applying to grammar, sentence, and term agreement. Using the frequency of terminology, a method can be created to link a knowledge

base, in the form of a dictionary of defined terms, to a text document. Using this understanding a computational method of extracting information from text can be created. This is a task with many different aspects to address such as what resources to use and what text corpus to construct for instance. Some of these issues will be discussed in greater detail in the following sections of this chapter.

The opinions and behaviour of other agents play an important role when making a decision or choice. This is particularly true when those decisions involve opportunity or valuable resources. Opinions expressed by word-of-mouth are a powerful form of information that influences our decisions and actions. Until recently, these opinions were not as easily expressed from one person to another. With the advent of the Internet, in particular the social web, ideas and opinions have never been so easily spread. Many tools now exist to quickly exchange and disseminate ideas with everyone connected to the World Wide Web. This information can encapsulate facts and opinions related to political campaigns, marketing campaigns, financial markets, products, places, and individual people. This information, often freely available, is largely unstructured and its interpretation has become of interest to the scientific community and business world alike [64] [80]. Making news and text information machine readable and interpretable has resulted in the fields of sentiment analysis, opinion mining, natural language processing, text analysis, and content analysis.

One task of text analysis is the classification of positive and negative opinions from text. Classifying reviews of products and movies has attained considerable attention [51] [100], where determining the polarity or sentiment of the reviews is the intended outcome. One particular critique of opinion mining for reviews is that the text is typically terse and discusses one topic, the particular movie or product in the text. In a news article, opinions can become mixed or may not even contain strong evocative sentiments. Issues such as whether a selection of text are topical or relevant is less ambiguous with product or movie reviews than with general news. These issues can be somewhat addressed by looking at the type of news article and sources being used in a collection of text. This question has been considered in studies where text and sentiment analysis programs have analysed a collection or corpus of text that is specific to a particular topic [28] [51] [55]. This is a concern as a document may contain off-topic content but contribute to the sentiment of the document.

Methods of text classification, popularised by the machine learning community, have been used extensively in statistical based text analysis. Many of the methods of machine learning require a corpus of annotated text that is used with a chosen algorithm or ensemble of algorithms that can learn the affect of a term or text document. Techniques such as Naïve Bayes, maximum entropy, support vector machines, latent semantic analysis, and latent dirichlet allocation, have shown to perform well in specific scenarios with advantages and disadvantages to each approach [14] [33] [49] [61] [75] [81]. Typically, classifiers depend on a large volume of text to act as a training set for the algorithm to be able to identify the necessary features. A dependence on a large volume of annotated text means the choice of text is important for a good training

set estimation. Typically classifiers will perform poorly outside of the context of the original training and testing set. Machine learning algorithms can perform well for specific tasks but may lack adaptability as supervision is needed to decide on the features and training sets. This remains a disadvantage to methods that rely on an annotated corpus and restricts the use to certain sources, while other sources may not readily have access to annotated data.

Concept based approaches to text analysis leverage ontologies or a knowledge base to perform analysis, example resources include WordNet [71], the affect related extension WordNet-Affect [93] and Sentiwordnet [37] have been popular choices in the literature. By using a knowledge base it may be possible to infer meaning and features associated with concepts contained in the text. Concept based methods rely on the syntax of text as they aim to detect sentiment that is more subtly expressed. The performance of this method relies greatly on the quality of the knowledge base. A detailed understanding of the concepts and knowledge of a particular domain and their relationships is required to allow the opinion mining system to perform adequately.

As new mediums and platforms of communication emerge, the availability and creation of data will increase, many using natural language as the intermediate form of communication. With the growth of social media and the web, filtering content, subjectivity detection, and in particular evaluating sources will become increasingly important and necessary for all forms of learning and analysis. More comprehensive knowledge bases will need to be combined with reasoning methods to ascertain better understanding. This will support a better understanding of natural language and opinions, turning unstructured data into machine readable formats, creating systems capable of handling complex knowledge tasks. New large scale computing paradigms will only aid these tasks in the future as better methods of aggregating and handling data will be combined with better resources and methods of learning.

2.2.1 Text Preprocessing

An analysis of a news article may reveal opinions, sentiment, or affect towards an entity or about a concept. In the work by Osgood et al. in *The Measurement of Meaning* [24], a concept is judged or measured based on a series of scales. By using these scales, a concept can be localised to a point of agreement, and as more than one dimension can exist, this point will be contained in what is described as a "semantic space". Within this semantic space, direction and distance from the origin represent quality and intensity of meaning. When referring to a term, the direction can relate to polarity and the distance as a measure of how extreme the term is. Osgood et al. identify three dimensions in particular that can be attributed to meaning namely activity, potency, and evaluation. From the evaluation scale, polarity can be expressed by negativity and positivity, and using this insight it may be possible to describe a concept as being overall negative or positive in dimension.

Typical methods of text analysis and opinion mining look at different levels of granularity, these include:

- Document level
- Sentence level
- Aspect Based
- Comparative sentiment analysis

Phrase level sentiment aims to diminish the problem of conflicting polarity on a document level. In the work by Turney [100], previously defined phrase chunks and terms are used to determine the opinion of a document. Sentence level analysis has also been examined by looking at the items in a fragment of text or sentence. Examining the subjectivity of sentences and phrases from an aggregate collection of opinions may better define the overall opinion of a document [85] [107]. Aspect level analysis consists of targeting certain entities and the polarity or sentiment related to them [51] [98].

Pang et al. originally looked at document level analysis and classified the overall polarity of the text [81]. This is a useful interpretation for movie and product reviews where the topic of the review is largely about the item in question [79]. The overall polarity of a document will likely be directly related to one topic. A document or news article by contrast may contain multiple opinions about various topics or entities and additional methods of segmentation or granularity of analysis, such as word level analysis, may be necessary to accurately filter and disambiguate opinions.

Looking at the word level of documents and locating key words in a text is very intuitive and accessible in implementation, and as such makes it a popular method of text analysis. Predefined affect categories in a dictionary such as positive, negative, or weak comprising of tagged terms or words are counted based on their presence in a piece of text and the affect or sentiment of the text is determined by the overall frequency of these categorised terms.

Word level analysis simplifies many of the assumptions of textual analysis. It is also a very transparent approach and economical in its process. Using keywords however does have some shortcomings and trade-offs as is the case for any approach taken. The reliance on affect or sentiment words being present only allows for a surface level interpretation. Many nuances of a language are missed and only the most obvious message is extracted. The underlying meaning can often be missed in certain cases where strong emotion is expressed but no keywords or affect terms are used. Additionally much of the analysis is dependent on the quality of the affect categories or dictionary being used. A lot of effort and research has been devoted to the generation and evaluation of dictionaries and lexicons using automated methods and legacy resources [48] [51] [77] [94] [101] [106] [108]. A difficulty comes in evaluating these resources however, as subjectivity still exists in the method. Many of these studies and approaches focus on the application of the automatically generated resources.

Efficient and effective methods of text processing and information retrieval are needed to manage the deluge of text and information now available. Identifying key terms is one of

the simplest and most effective methods of processing large volumes of information. Before performing even this level of processing, a data structure is needed to represent terms and their occurrence in text. This is apparent in nearly all the text analysis studies presented in this section. The construction of a vector space model is one data structure that can represent term occurrence in text. To conclude this section a brief summary and motivation for this representation and its relation to the bag-of-words model is described.

The vector space model (VSM) has many appealing properties despite its simplicity. The extraction of information from text requires far less effort than other approaches of text and semantic analysis. The VSM model is a widely used representative model involving tasks such as word similarity, query and document similarity, and semantic relatedness [69] [101] [102]. It also forms the basis for the bag-of-words model used frequently as a starting point in text analysis methods.

The *bag-of-words* model was first referenced by the linguist Zellig Harris in *Distributional Structure* [46]. The model begins with calculating the frequency of occurrence of each word without noting position, grammatical relevance, or part-of-speech initially. Using the frequency of these terms as features of the space, either a document or total words in all articles released in a day, the vector space model can be constructed and provides the basis for much of the text analysis work presented in this thesis.

In this thesis, the elements of the VSM that will be used relate to the frequency occurrence of a given word. It has been stated that a lexicon will represent the terms in question, categorised accordingly, with their frequency in either a document or during a specific time period being computed. The frequency of occurrence in all documents in a day is the interpretation that is used in this thesis and forms the time series of term frequency discussed in more details in Chapter 3. Due to the emphasis on term frequency in the work presented here, the specific VSM used will be referred to as a term vector. Representing text in this way is the starting point of text processing and then allows more complex methods of text analysis to be performed and built on this representation such as content analysis.

2.2.2 Content Analysis

Content analysis was considered a broad area of research where behaviour was examined by scholars in the areas of political science, law, anthropology, and history. The pioneering work of Harold Lasswell saw general categories for terms being defined and the development of quantitative indicators for such [57]. This formalism and move towards quantitative efforts meant the methods of content analysis could benefit from computational power.

The General Inquirer (GI) [92] system was built on the foundational work of Lasswell and others to produce a content analysis system employing dictionaries of categorised terms to measure affect and sentiment in content among other concepts. The system merged two dictionaries of terms the Harvard-IV-4 psychosocial dictionary and Lasswell dictionary [92]. The merged

dictionary formed the (original) 182 categories that were to be the basis of the GI dictionary used in the system. The GI system allowed users to count affect-laden words in text. From this it was then possible to create a quantitative measure of affect in a large volume of text more easily.

Content analysis and machine learning approaches have been the foundation for several systems that have been used widely and across different disciplines to perform a number of text analysis tasks. Many libraries exist that perform typical text processing tasks and classification. *Bow* [70], and its accompanying front end "Rainbow", is one such system. This toolkit allows users to perform typical text preprocessing operations such as tokenisation, generate document and word vectors, make queries, perform Naïve Bayes classification, and calculate term frequency-inverse document frequency (tf-idf) scores. *GATE* (A general architecture for text engineering) is another open source system that performs text analysis and ontological investigation [25]. Modules in this system perform text analysis and classification with methods for part-of-speech tagging, ontology editing, and parse tree visualisation among others. A popular and widely used Python based platform for working with language and text data is the Natural Language Toolkit (NLTK). It contains interfaces to many lexical resources and known corpora in the text analysis community and has functions to perform classification, tokenisation, stemming, tagging, parsing, and semantic reasoning [10]. Although the UI is not as fully functional as the GI program or GATE, the API has provided a strong platform for commercial use and for the open source community to develop powerful applications.

Many studies have employed commercial and some publicly available machine learning systems in quantifying textual data and contribute to the literature on using text analysis systems with statistical and econometric models for financial data. One program in particular is the *Thomson Reuters News Analytics* (TRNA) system that acts as black box supplying quantified text and metadata relating to certain topics, events, or subjects to a user. Studies by Leinweber [59] and Mitra [72] have used the TRNA.

Systems have also incorporated non-dictionary based approaches with term frequency again acting as the feature extracted from text. An example of such a system is *DICTION* [47], which has been used to perform content analysis in finance [32]. A forthcoming study by Loughran and McDonald highlighted issues related to word misclassification with using DICTION for domain specific text [68].

The addition of a machine-readable lexical resource, in this case a "dictionary" of categorised terms, is necessary for the dictionary based approach to content analysis. Terminology is influenced by the topic or subject of discourse. The analysis of special subject languages by linguists has highlighted differences between term use in general and domain text. Computational linguistics has pushed terminology theory and analysis to being more application driven. By the observation of such applied work, new theoretical insights can be gained as well as the applied benefit. Terminology is related to the concepts that it is used to describe, and can represent the knowledge of a domain and stored as a lexicon [86]. The use of terminology to measure

qualitative information in text is used in this thesis to measure the tone of a text, creating a quantitative measure that can be used in statistical modelling to represent the text content.

Dictionary based content analysis relies on a given terminology categorised accordingly into a dictionary. One notable lexical resource is the GI dictionary previously described in this section. Some studies have looked at automatically generating these dictionaries and lexical resources. Turney has demonstrated methods of detecting semantic orientation of words by association with seed words determined to be definitely negative and positive [100]. Seed words have been used to expand a dictionary by associating additional words through semantic orientation, words that appear together in a similar context in text. This approach has been used with adjectives that acted as indicators of semantic orientation [48] [51] [94] [101] [106]. This method however still relies on some form of supervision by providing pre-tagged or pre-determined seed words.

Loughran and McDonald highlighted the potential for misclassification of terms using a word list that was more typical of general language text than domain specific text such as business news [66]. This conjecture was also considered in Li [62]. To address this potential issue Loughran and McDonald created a list of financial negative and positive words by examining the risk factors section of Form 10-Ks which reports possible failures and risks for a company. This word list has also been used in several studies where textual sentiment in finance and business was examined [42] [52] [65] [67]. These word lists were created by hand by the authors, include multiple tenses for words and have not been compiled in a way that would be consistent with a linguists interpretation of a lexicon of terms. These differences mean this word list and others that have been compiled in financial literature without references to terminology construction, such as in Henry [50], may not have undergone as much scrutiny as a formally constructed dictionary such as the one included in the GI program.

A lexicon is an important resource for content analysis. Often they are created for a specific purpose. The GI dictionary was constructed from the original dictionaries by a team of researchers looking at the frequency of words in general language text and determining which words have polarity and would be regarded as terms. Terms were chosen to encompass and represent a concept rather than just one simple meaning. Whether a general purpose lexicon such as this is useful for a specific domain application has seen less attention in content analysis literature. Specialist and domain languages contain a variety of "named entities" and words that can act as key terms for a specialist domain. Often these terms can form a signature of the subject domain [1]. It has also been noted that "special languages have a higher rate of repetition of lexical items than general language texts" [87]. This may indicate that the frequency of the domain specific named entities can carry knowledge and meaning that may be valuable. Choosing a lexicon to best represent the specialist domain would ideally be done by using expert knowledge and opinion of the area. For domain news the frequency of terminology can be observed and used to filter for "good" terms or named entities. Published definitions of terms or glossaries of specialist words may assist in capturing the typical language of a domain. This is the main question asked by Loughran and McDonald [66] who show that word lists and

legacy dictionaries, such as the GI dictionary, may give different interpretations for domain text than for general language text. This leads to the question of whether including knowledge about domain terms and words can help in summarising affect and sentiment in a specialist language corpus.

The use of a domain word list with a general language dictionary to improve the content analysis method of specialist language text is a question that is investigated in this thesis. In the domain of finance and accounting, researchers have focused mainly on the negative and positive categories but in the vast majority of studies have not attempted to incorporate knowledge about the domain. The case studies presented in this thesis use the system developed by first using the GI dictionary as the basis of the content analysis approach, identifying sentiment in finance text using the negative category of the GI dictionary. Then the GI dictionary acting as a base dictionary is combined with a domain glossary of words to add more information about the context of certain GI terms for a domain text. The assumption is that by analysing domain text and including a glossary of words or terms about that domain from an expert or reputable source the count of affect frequency from the GI dictionary can be made more accurate.

Content analysis systems have been used to measure the tone of a text and this output has been used in inferential and predictive models for a number of uses, more notably to track the impact of news and improve price predictions of financial assets. Many of the studies reviewed here have drawn on methods of text analysis and econometrics separately and performed analysis in isolation. Fewer studies have attempted to combine these methods into a single system of analysis. Even fewer have attempted to generalise this implementation to allow analysis of different text sources, types, and financial data. It is the intention here to combine these functions into a system that can take any number of text and time series data input to analyse the impact of news on financial assets.

2.2.3 Constructing a Domain Corpus

A number of texts documents collected and organised in a systematic way is considered to be a corpus. When building a corpus of text, consideration must be given to developing a balanced, representative collection of information regarding a specialist area or topic. The British National Corpus (BNC) is a widely known and used corpus that is representative of contemporary English and is still primarily concerned with well-known authoritative publishers [1]. Statistical methods of text analysis and natural language processing start with building a corpus of text from known sources and structuring the text in such a way that it is machine readable. Additionally, an important aspect of this study and intended application of text analysis is the “time” element of the news or text. This is included in the machine-readable metadata when structuring the corpus documents. The date of publication of each text is important as a time series variable is constructed by the system presented in this thesis from the text corpus.

The information extracted from text can vary greatly depending on the source and type of

text used. It becomes difficult to find or construct a corpus of text that is representative of a specific topic. For instance, it can be assumed that the language used in research papers and journals will have terms that ordinarily will not occur in general language text with any large frequency [1]. However specialist terms may be widespread in the domain text. The approach of Manning in *Foundations of statistical natural language processing* [69] is that having more text for the purpose of training is more important than concerns of corpus balance, meaning using all the available text is often beneficial. Alternatively, applications in the area of finance have shown that by choosing authoritative sources, and relevant articles, the results will be more in tune with intuitive beliefs and also theoretical models of behaviour [42] [96]. The construction of language resources such as the British National Corpus (BNC) have also shown that corpus linguists prefer reputable sources [1]. Several studies have shown that topic relevant news about assets, companies, and financial markets can also influence predictability [17] [20] [80]. By employing computational methods to collect a large volume of text from a reputable source and organise the corpus in a diachronic manner these concerns can be addressed.

Previous studies in finance have focused their efforts on authoritative news publications, company filings and disclosures, internet messages, and in recent time's social media. Social media in particular has acted as a new source of public information that has seen huge growth in the quantity of information being produced. Authors have sought to extract different meanings from the different information sources. In the domain of finance, authors have attempted to identify sentiment regarding market participants, analyst opinions, and public sentiment for markets, individual institutions and even economies. Company filings, annual reports, and earnings announcements have been used to examine sentiment expressed about a company [29] [52] [63] [66]. For corporate filings and earnings announcements, the information contained in these reports discusses the firm's performance, its place in respective markets, and additional comments, which might reflect the sentiment of relevant events that would be of interest to the company. Corporate announcements and statements are usually released quarterly or on an annual basis. The lower frequency of announcements means a specific approach must be taken to assess the impact around the time of the release such as performing an event based study.

The use of informal sources of online messages and social media has also been prevalent in recent studies. The content and volume of messages boards have been shown to proxy for market behaviour [8] [27]. More recently with the advent of social media, great interest and emphasis has been put on looking at fluctuations, volume, and analysing content of social media and blogs [15] [76] [110]. Social media and many forms of online media tend to be unedited and unregulated. Due to this, the text data tends to be noisier, requires more effort in evaluating the messages and is harder to extract meaning from overall [56]. In the case of each source, strategies for monitoring, collecting, and understanding the nuances of the platform need to be considered.

News articles are typically published every day and are well structured typically being contained within a particular section or column of a newspaper or site and organised by topic.

Daily news articles from authoritative news sources have shown to give reliable results in financial literature when combined with content analysis based techniques to examine the text content [8] [26] [42] [96]. News and messages published sporadically throughout the day but are typically published every day. By aggregating all text published in a day to equal one daily observation, a daily time series of news sentiment can be constructed that is more consistent and evenly sampled in the number of observations in time than at higher time frequencies where the time of publication is less consistent. This is particularly beneficial for the regression based studies in the literature where consistent well-formed daily time series of news are aligned with daily price series allowing more parsimonious models to be estimated and relying on fewer assumptions for the model and data. Many of the studies in financial literature that incorporate text analysis methods use daily time series.

2.3 Data Modelling and Statistical Estimation

The text analysis methods used in the literature, particularly those in the domain of finance, have been presented in Section 2.2.2. Their relative advantages and disadvantages have been discussed as have issues surrounding corpus construction and using lexical resources. Using the results produced by the text analysis systems and methods it becomes the task to see how these new data can be related to financial variables and test hypothesis about the impact of sentiment on financial markets. The range of time series and econometric models is wide. In the following section time series models that have been used frequently in the literature are discussed and how they handle the idiosyncrasies of financial time series data. The objective is to use these models to combine financial variables with the output from the content analysis method of analysing text to relate the effects of news with financial markets.

2.3.1 Linear Regression

Linear models attempt to quantify the relationship between present values of a random variable with the present value given all available information. In financial time series analysis this random variable can include the returns or price of an asset. Considering the return of an asset r_t, a simple model will attempt to explain or capture the linear relationship of r_t using information available before time t. This information typically begins with historical values of r_t (linear regression) and can include additional random variables with useful information (multiple linear regression). Additional variables might contain information about the economic environment and market structure in which the assets value is determined. How these variables relate to each other plays an important role in model specification. Understanding how correlation and autocorrelation occurs is a basic tool for studying time series and as such is the building block of linear estimation for stationary series.

Often a marginally statistically significant autocorrelation is observed in financial returns

[95]. This suggests that a previous return value r_{t-1} may be useful in providing information about r_t. This can be summarised in a typical autoregressive equation:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \varepsilon_t \tag{2.1}$$

where

ϕ_0 = the intercept value
ϕ_1 = the regression coefficient of the previous return value
ε_t = the residual term assumed to have constant mean and variance

The model shown in Equation 2.1, demonstrates an autoregressive process of order one, which will be abbreviated as AR(1). The AR model is similar to the widely used simple linear regression model and is conditional on past observations, in this case returns. The AR model is used in financial literature to measure the persistence of dependence in a time series. Due to being weakly stationary, the mean and variance are finite and time invariant, and $|\phi_1| < 1$ is taken to be true for an AR series, ϕ_1 measures this dependence. In the case of a simple linear regression model, the dependent coefficient value can assume any fixed real number.

For a model that is determined to be a good fit, the residuals will behave similarly to white noise. If some estimated coefficients are not very different from zero, then simplification of the model and removal of parameters may increase the accuracy of estimation. If the residuals of the autoregressive model exhibit serial correlation, then increasing the order of the model may help reduce this correlation in the residuals, increasing the accuracy of the estimation. Linear regression models have been the basis and starting point for a number of studies incorporating sentiment as a variable into an econometric model, these studies are discussed in more detail in Section 2.4.

2.3.2 Autoregression and Heteroskedasticity

The dependent variable of interest studied in the evaluation in Chapter 4 is the return of a financial asset. Variability in time can be seen in the returns series and summarised by calculating the variance and standard deviation of the series. Changes in the scale of the variance parameter over time, indicates heteroskedasticity in the series. In the pioneering work by Engle [36], a stochastic process is used to represent this variance in returns and is defined as having an unconditional mean of zero and conditional variance as a linear function of previous squared residuals from the model fit, the so-called autoregressive conditional heteroskedasticity (ARCH) model. The variance of returns is then conditional and dependent on information contained in previous returns.

The model shown in Equation 2.2 is a type of basic linear regression with dependent variable r_t and independent variable r_{t-1} demonstrating an autoregressive (AR) process. This representation, AR(1), determines r_t to be centred around $\phi_0 + \phi r_{t-1}$ conditional on past returns, where

r_t is not correlated with r_{t-i} for $i > 1$. In real world cases the conditional expectation of r_t is not determined exclusively by the previous period return. We can generalise the basic AR model to include additional lags up to order p giving AR(p) (Equation 2.2).

$$r_t = \phi_0 + \phi_1 r_{t-1} + \cdots \phi_p r_{t-p} + \varepsilon_t \tag{2.2}$$

where:

ϕ_0 = the intercept value
ϕ_p = the regression coefficient of the previous t-p return value
ε_t = the residual term assumed to have constant mean and variance

Previous returns up to an order p now jointly determine the expectation of r_t. This is the same approximation as a multiple linear regression model with lagged values of the independent variable being included in the model specification. The choice and order of p has to be determined either empirically, using the autocorrelation function (ACF) and partial autocorrelation function (PACF) or by using some information criteria such as the Akaike information criterion (AIC) [3]. There has been much research in the area of order determination, often using the autocorrelation function or an information criteria. To use the partial autocorrelation function (PACF) for an AR(p) series, the lags up to a statistically significant correlation period p are chosen. Where the PACF is zero, at lag p for instance, the cut off for an AR process can be found. Information criteria are also available to determine the lag order, the AIC is defined as:

$$AIC(\ell) = \ln(\tilde{\sigma}_\ell^2) + \frac{2\ell}{T} \tag{2.3}$$

where:

$\tilde{\sigma}_\ell^2$ = the maximum-likelihood estimate of the variance
T = the sample size
ℓ = the lag order

The second term of the AIC equation acts as a penalty function for the number of parameters used. The AIC measures the goodness of fit of the model, taking the number of variables into account and estimated using the maximum likelihood. The Schwarz-Bayesian criterion (BIC) is another commonly used and reported statistic. BIC will often favour a shorter order than AIC but there is not strong evidence to suggest one measure is superior to another in application. Due to this, knowledge of the data being modelled and the problem being addressed, combined with the simple assumptions that result from this knowledge are the dominant factor in choosing lag order, this is particularly true of AR components [99].

2.3.3 VAR and Multivariate Modelling

Multivariate time series analysis involves methods for examining multiple series and is used extensively in econometrics and statistical analysis. A typical definition of a multivariate time series is one that is composed of multiple single series, these can be referred to as components. This leads to a representation of a number of series under one matrix. For autoregression based models, n number of lags of a series regressed upon itself are included, this is the basis for the vector autoregression (VAR) model [90]. The VAR model is a flexible and easy to use model for analysing multivariate time series. It is particularly useful for describing the dynamic behaviour between time series and frequently used in forecasting. A VAR(p) model often has multiple parameters resulting in many interactions and feedback between variables.

When considering two variables y_t and z_t if it is believed that one variable will be affected by past and current observations of the other variable a bivariate system (in the case of two variables) can be formed (Equation 2.4 and Equation 2.5) that will constitute a first-order vector autoregression model:

$$y_t = \beta_{10} - \beta_{12} z_t + \gamma_{11} y_{t-1} + \gamma_{12} z_{t-1} + \varepsilon_{yt} \tag{2.4}$$

$$z_t = \beta_{20} - \beta_{21} y_t + \gamma_{21} y_{t-1} + \gamma_{22} z_{t-1} + \varepsilon_{zt} \tag{2.5}$$

Where y_t and z_t are stationary and ε_{yt}, ε_{zt} approximate to white noise and are uncorrelated.

The system described by Equation 2.4 and Equation 2.5 allows the variables to affect each other. This system of equations can be transformed into a more compact form to obtain the standard form of the vector autoregression, Equation 2.7, derived from the definition in Equation 2.6.

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \epsilon_t \tag{2.6}$$

where:

$$B = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}$$

$$\mathrm{x}_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix}$$

$$\Gamma_0 = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix}$$

$$\Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

The standard form of the VAR model then becomes:

$$x_t = A_0 + A_1 x_{t-1} + e_t \tag{2.7}$$

where:

$A_0 = \mathrm{B}^{-1}\Gamma_0$
$A_1 = \mathrm{B}^{-1}\Gamma_1$
$e_t = \mathrm{B}^{-1}\epsilon_t$

The multivariate generalisation of the VAR model can finally be described as:

$$x_t = A_0 + A_1 x_{t-1} + A_2 x_{t-2} + \ldots + A_p x_{t-p} + e_t \tag{2.8}$$

where:

x_t = an (nx1) vector containing each of the n variables included in the VAR
A_0 = (nx1) vector of intercept terms
A_p = (nxn) matrix of coefficients
e_t = (nx1) vector of error terms

The aim of this method is to create a parsimonious model [90]. Sims' approach to the multivariate VAR requires the inclusion of appropriate variables in the model and to choose a lag length [90]. The ultimate goal of this model is to uncover the inter-relationships between variables in the model rather than make short-term forecasts [34]. As the VAR Equation 2.8 contains predetermined variables and the error term is serially uncorrelated with constant variance, each equation in the VAR model can be estimated using ordinary least squares, as stated in Enders [34]. Tetlock in [96] specifies a VAR model incorporating sentiment and DJIA returns to observe any relationships between variables and uses ordinary least squares to estimate the equations as it is stated that this estimation provides consistent and asymptotically efficient estimates of the A matrix of coefficients as the right-hand-side variables are predetermined and are the same in each equation of the model. The VAR model is one of the models included in the modelling component of the system developed in this thesis as described in Chapter 3. The prevalent use of the VAR model in the literature, it's ability to determine inter-relationships with other variables included in the model, and it's relative ease of estimation make it a good choice for modelling the impact of sentiment on financial time series.

2.3.4 Volatility Modelling

The generalized autoregressive conditional heteroscedasticity (GARCH) model has been one of the most influential models in volatility analysis [16], providing a reliable method of computationally calculating volatility. With a particularly easy and robust method of estimation and the intuitive assumptions governing its construction, the GARCH model presents itself as a good choice for investigating the effects of volatility.

To construct the GARCH model a stochastic process is specified with zero mean and conditional variance. The conditional variance is specified by a linear function of previous squared residual of the stochastic process following an autoregressive process. If the stochastic process are financial returns then from the distribution of returns for a period t, a constant mean and time varying conditional variance can be computed. A lagged variance term can then be included, this generalises the mode giving the GARCH estimation. This model of volatility has proven to be difficult to outperform in many instances [95].

The definition of the GARCH model begins with the distribution of returns, conditional on previous observations for period t defined as:

$$r_t \mid r_{t-1}, r_{t-2}, \ldots \sim N(\mu, h_t) \tag{2.9}$$

where μ can be assumed to be constant and conditional variance h_t is specified as:

$$h_t = \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1} \tag{2.10}$$

where:

h_t = the conditional variance
α = the coefficient for the return residual
r_{t-1} = the previous return value
μ = the mean of returns
β = coefficient for the previous variance

The four parameters are constrained as $\omega \geqslant 0$, $\alpha \geqslant 0$, $\beta \geqslant 0$. The GARCH model relies on one previous squared residual and conditional variance term, after the initial stage the model can be solved recursively. The model is stationary and well specified if $\alpha + \beta < 1$, which represents persistence, the rate at which a volatility effect will decay.

Several studies have looked at the effect that proxies for sentiment and information have on volatility. These studies looked at the effect that the arrival of information has on volatility [4], incorporating a sentiment index into the GARCH estimation directly [12], and using regression based analysis to relate text sentiment with volatility [8]. The GARCH model is used in the system implemented here to calculate volatility for the returns series (Chapter 3 Section 3.4.2). This estimation of volatility is used in the evaluation and testing of the system to assess any relationships with the sentiment variable.

2.4 Information and Financial Markets

The finance community has expressed increasing interest in sentiment analysis and text analysis methods in recent times as evident from the studies presented in this chapter thus far. Many studies in the finance community incorporate different theories and assumptions about

the impact of sentiment or its role in financial markets. Traditional beliefs held in finance of an efficient market with rational participants are not as definitive as first theorised. The strict idea of an efficient market, that all information is incorporated into price at all times, would lead investors to believe that news and exogenous information would have no impact on the markets, having already absorbed any and all possible innovations. The existence of speculative bubbles however, suggests that market participants are acting in a way that is devoid of information about the underlying value, exhibiting irrationality and making decisions based on emotional bias or assumptions and possibly overreacting to new information [89].

Financial analysts, particularly those engaged in fundamental analysis employ a number of data sources to make judgements on the true value of an asset. Data sources can include raw financial data such as price, exchange rates, or company earnings, market and industry reports or announcements, emergency and disaster events, geopolitical events, and government regulation. This information can be unstructured, qualitative in nature, and difficult to aggregate. Much of this information is reported in news and contained in text. As such, quantitative methods to measure and aggregate this qualitative information are appealing.

Information contained in news may influence market participants. This information may then provide some predictive insight into the reaction of investors and therefore market movements. The idea that there exists market participants trading on information, and those who trade on misinformation or stale information, described as noise, was discussed in the formative work by Fischer Black in *Noise* [11]. Black also highlights that noise is necessary for a market. Although noise may contribute to making a financial market imperfect, it supplies liquidity, of which those trading on information can take advantage of. As the number of noise traders trading increases, those trading on information must increase their position, taking on more risk. For liquidity to exist in a market, prices will become less efficient and as stated "markets are efficient almost all of the time" partly due to this. Black's conjecture that investors with no access to insider information or who are misinformed for numerous reasons such as stale information or make decisions based on noise thinking it to be valid information. These market participants act irrationally with opposing traders being the rational traders using valid information.

The presence of noise traders in financial markets, although an accepted theory, has been somewhat overlooked when considering price formation and value estimation. Fama [38] has argued that irrational investors are met by rational investors in a market taking positions against them and correcting price, pushing it back to expected fundamental values. It is thus believed that noise traders cannot affect price too much and their effect will be transitory. The assumptions underlying the rational traders ability to exploit noise traders ignores the limits of arbitrage somewhat. Arbitragers are likely to have short horizons for trading and are somewhat risk averse. As such they are limited in the size of position they may take against noise traders and in liquidating the mispriced asset [30]. Noise trader risk is difficult to specify directly, however De long et al. [30] note that by looking at the effect of unpredictability of irrational behaviour on rational investors much can be learned about the impact and maybe facilitate

better risk management.

Tetlock makes the statement that despite the theory behind what sentiment in text represents, either based on liquidity and trading effects, or noise trader behaviour, the variable extracted from text can act as a proxy for anomalies such as investor sentiment or risk aversion [96]. However the interpretation, be it noise traders, effects of an inefficient market or the idiosyncrasies of trading, Tetlock demonstrates that information extracted from the content of news can have a statistically significant impact on the price of assets. The behavioural finance theories behind what the proxy for sentiment in text represents however provide interesting insights into what type of information is begin captured by using content analysis to summarise text into a quantitative value.

The use of proxies to represent information that is difficult to measure is not a new concept in financial or statistical analysis. One prominent example in finance is the US-based Michigan Consumer Sentiment Index (MSCI), which is a monthly report quoted extensively in U.S. financial and mainstream news media. A survey is issued with questions concerning personal finance, business conditions and purchasing power. The answers to the survey are collected and normalised to create a series[1]. Through the MSCI, consumers express their anticipation to changes in the economy by their personal opinion and sentiment. The MSCI then acts as a proxy for consumer sentiment. Some studies have stated that the results of the MSCI serve as a better forecast of US inflation rates when compared with complex econometric models [7]. Studies such as these show the benefit that proxy for hard to measure information can have.

Another example of a widely known and used proxy in finance is the VIX index, referred to as the "Index of fear"[2] [105] and can act as a proxy for market volatility. Evidence has been presented about the useful forecasting ability of the VIX series [13]. It has been used in conjunction with other market indicators as a proxy for investor sentiment predicting possible reversals in the financial market.

Studies in finance have used different means to generate proxies for investor sentiment and alternative information, this has included the use of surveys, trading volume, order flow, dividend premiums, and implied volatility of options amongst others [9] with more recent studies relying on computational methods to analyse news and text content.

Several studies in finance have used proxies extracted from text data and news using text and content analysis methods. The content and activity of internet messages boards *Yahoo! Finance* and *Raging Bull* were investigated by Antweiler and Frank [8]. The objective was to see if useful financial information was contained in the message boards. The study addressed the predictability of returns and volatility using different interpretations of message content and the volume of messages. The study looks at the volume of messages posted and message *bullishness* (ratio of messages classified as buy or sell using a Naïve Bayes classifier) for their ability to predict returns. The financial returns examined are for 45 stocks that appear in the Dow Jones Industrial

[1] http://data.sca.isr.umich.edu/fetchdoc.php?docid=48865

[2] http://www.cboe.com/micro/vix/vixwhite.pdf

Average and an exchange-traded fund that mimics the S&P500. They employ correlation, contemporaneous regression, and panel regression models to determine the impact that news media has on stock returns and volatility. Their findings suggest that an increase in message volume could predict negative returns on tomorrow's prices while an increase in bullishness is associated with an increase in contemporaneous returns but statistically insignificant in other cases. For message volume on stock returns the results suggest a unit increase (100% increase or double the number of messages) in the number of messages will result in a 0.2% decrease in stock price. This price change is small and it is noted that transactions costs would diminish the potential economic gain. Agreement in messages and any association with greater trade volume is also investigated. It is found that by conducting a contemporaneous regression, disagreement occurs with higher trading volume. This result supports theories of disagreement between market participants inducing trading, although the authors also note a reversal next day, the effect of message board activity predicting volume is a more significant effect than the reversal. Relating message posting volume to volatility is also examined. Measures of volatility used include realised volatility [6], GARCH [16], and Glosten-Jagannathan-Runkle (GJR) GARCH [44]. A relationship between message posting and volatility is found but the inclusion of trading volume sees a reduction in the predictive power of message volume. Despite this potential multicollinearity of variables, the effect of message board volume is not diminished. The disagreement factor's effect is seen to be weak. This study demonstrates how online text and news can contain useful information in explaining changes in returns.

In Tetlock a study is undertaken where the content of news is used to predict movements in stock market activity [96]. The focus of the study is on the content of the *Abreast of the Market* column published daily in the Wall Street Journal (WSJ) and its ability to predict movement of the stock market index the Dow Jones Industrial Average (DJIA). Using the General Inquirer (GI) content analysis system, the content in the WSJ column is measured and a daily time series that acts as a proxy for investor sentiment is constructed. A pessimism metric, which relates to negative and pessimistic words in the column, is created by using principal components analysis to reduce the number of dimensions of the categories in the GI system occurring in the column. The negative and weak categories are also used in the analysis of the column as they occur in the GI system. Using a vector autoregression model, Tetlock shows the explanatory power the negative sentiment metric has on the movement of the returns of the Dow Jones Industrial Average.

One of the main results that Tetlock shows is that an increase in media pessimism robustly predicts a decrease of 8.1 basis points (1 basis point is equal to 0.01% change in price) in DJIA returns, with 4.4 and 6 basis points for the negative and weak categories respectively. This effect is then reversed over a trading week. High market trading volume, as measured by the New York Stock Exchange (NYSE), is predicted when pessimism is low or high. One of the final points highlighted by Tetlock is that low market returns predict high pessimism. Tetlock states that these findings support the idea that content may act as a proxy for investor sentiment. The

reversal of the effect of media pessimism and lack of long term effects suggest that the effect is not based on information or awareness of fundamental valuation. However despite previous studies, the ability of the pessimism metric to predict volatility is weak. This study demonstrates the impact that negative news can have on a major market index and how this effect can be represented using a proxy for the content of the text.

The influence of news text in financial markets is prevalent and the impact can be deciphered using the volume and content of news. In the work by Garcia [42], support is given to the results presented by Tetlock [96]. The study conducted by Garcia attempted to extend and verify previous claims of the impact of investor sentiment detected using a proxy from news computed using content analysis. This was done by using a corpus of news from the New York Times from 1905 to 2005, a substantially large increase in the sample size from previous studies. The corpus was built using two columns the *Financial Market* and *Topics on Wall Street* column which discussed stock market performance and industry news. A novel contribution of the study by Garcia is the use of the NBER business cycles data to proxy for recessionary periods in the economy. The media variables constructed consist of the counts of the financial negative and positive word lists from Loughran and McDonald [66] in each column. An OLS regression model is built with dummy variables to account for business cycles and control variables for time series anomalies. The main findings reiterate that of previous studies that news text helps predict market returns (DJIA), while during recessionary periods the impact of the negative sentiment is stronger but less significant than in expansionary periods. Garcia finds that by using the negative sentiment series extracted from the NYT columns, a one standard deviation change in negative words results in a 12 basis point impact on the DJIA during recessionary periods and 3.5 basis points during expansionary periods. In the study by Tetlock the time varying impact of sentiment is also noted but not as thoroughly investigated. A partial reversal is also noted which reinforces the idea of the impact being a non-informational event but that of investor sentiment and a reaction to news.

The conclusion drawn is that a proxy for information contained in news may have explanatory power for financial returns. As financial markets are dynamic demonstrating significant changes in the distribution parameters of financial data, it may be assumed that sentiment in some form plays a role. How to analyse and compute a proxy for text then becomes the concern for an investigator. Creating and implementation a system to collect, extract, measure sentiment and determine when and to what degree this proxy for information impacts returns is the central theme of the work presented in this thesis.

2.4.1 Data Processing in Finance

A wealth of data is becoming available with databases of financial data being released for free and an even larger volume of unstructured data being published on the internet. Although these data are available, issues can arise with the collection and aggregation of such information.

Problems such as conflicting, non-unique, and ambiguous variables surround data collection and definitions. Time zones, price quotes, and non-synchronous data can mean different time series of price data being retrieved from exchanges and brokers. It may not always be possible to accurately define and sample data but a good knowledge and description of the dataset is beneficial and contributes to competent model estimation. These are some of the considerations in data collection and aggregation.

Synchronisation is another issue in creating an adequate dataset. Many variables will be recorded according to different frequencies of observation. It is often not possible to interpolate these data. Seasonal effects and other time series components must also be considered when preprocessing has been performed. Collecting data from different sources is often a difficult task that burdens many studying in the area of econometrics and finance. Databases and vendors aim to relieve this fundamental problem of data curation. Platforms such as Tradestation and Bloomberg create standards to deliver data in raw formats that are clear and concise. In recent times, free online sources such as Quandl[3] supply data in a standardised format from freely available authoritative sources such as the National Bureau of Economic Research (NBER), U.S. Energy Information Administration (EIA), and multiple exchanges worldwide. Automated systems can benefit greatly from these resources, integrating with data curation methods. It is a priority of such platforms to allow the easy and fast dissemination of data. Data processing methods implemented in the system in this thesis rely primarily on the Quandl database as the main source of financial data. The Quandl API allows easy integration with the statistical modelling component that will be described in more detail in Chapter 3.

Data from different sources can be difficult to aggregate, often due to formatting and coding errors. Specific platforms may have conventions in place for data, but aggregation and scaling may still be an issue. Arranging data in a uniform way is often the task of an external system or program. Statistical languages such as *R* are widely used and often the basis for developing statistical software for computing and visualisations. Complete systems such as *RATS* (Regression Analysis of Time Series), *SPSS* (Statistical Package for the Social Sciences), and *Stata* have also solved this issue without traditional programming experience. In recent times, the increasingly popular *Pandas*, a Python based software library built on top of the low level package *NumPy*, has provided very rich time series functionality. These systems and platforms solve the problem of data aggregation and the development of complex data structures to represent quantitative data with methods of alignment, grouping, merging, and additional manipulation. The chosen language for performing the statistical analysis in this thesis is *R* as there exists a wide variety of statistical libraries and visualisations. It is freely available with a general public license and has been used as the statistical engine integrated into many open source and commercial applications. For these reasons and its efficiency as a statistical language it was chosen for the implementation for the statistical methods and time series modelling in this thesis.

[3] https://www.quandl.com/

Table 2.1: A number of studies that have combined analysis of financial assets and text analysis methods. All studies use a bag-of-words approach to content analysis except those noted (*) which include triplets, OpinionFinder, classification, and Latent Dirichlet Allocation respectively to construct features.

Author	Text type	Text Source	Items
Wuthrich et al. (1998) [109]	General news	Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg	-
Peramunetilleke, Wong (2002) [82]	Financial news	Olsen Data	40 headlines per hour
Fung et al. (2003) [41]	Company news	Reuters Market 3000 Extra	600000 news articles
Antweiler, Frank (2004) [8]	Online messages	Yahoo! Finance, Raging Bull, Wall Street Journal	1.5 million messages
Henry (2006) [50]	Earnings press release	Compustat database	1366 annual press releases
Das, Chen* (2007) [27]	Online Messages	Message boards	145,110 messages
Tetlock (2007) [96]	Editorial Column	Wall Street Journal	4000 days with news
Tetlock et al. (2008) [97]	Financial news	Wall Street Journal, Dow Jones News Service	350,000 stories
Li (2010) [62]	10-K and 10-Q filings	EDGAR	140,000 10-Q and K filings
Loughran, McDonald (2011) [66]	10-K annual report	EDGAR	50115 10-Ks
Bollen et al.* (2011) [15]	Tweets	Twitter	9853498 messages
Engelberg (2012) [35]	Wall Street Journal	Dow Jones News Service	1888868 observations
Schumaker et al.* (2012) [88]	Financial news	Yahoo! Finance	2802 news articles
Jegadeesh, Wu (2012) [52]	10-K annual report	EDGAR	45860 reports
Yu et al. (2013) [110]	Social Media, Blogs	Twitter, Google	52,746 messages
Jin et al.* (2013) [53]	General news	Bloomberg	361782 news articles
Garcia (2013) [42]	Market News	New York Times	27449 days with news
Liu, McConnell (2013) [65]	Firm Specific News Stories	Dow Jones News Service	1009 acquisition announcements

2.4.2 Text and Data Processing Systems in Finance

The use and creation of systems for the analysis of text and time series data in financial and computational literature has gained momentum in recent times. One of the present challenges is in combining econometric methods with that of content and sentiment analysis. Following from financial and behavioural theories of sentiment in markets the question is not whether investor sentiment plays a role in asset prices but how to measure it and incorporate this measure into additional information models [9]. Whether news media or information extracted from such is or is not a proxy for investor sentiment, noise trader behaviour, or market microstructure effects, it is still possible to obtain useful information contained in news that can be quantified and help inform decision making by explaining movements in financial markets.

Table 2.1 shows a number of studies that have used text analysis methods with statistical methods to assess the impact of information on financial instruments and indices. The studies presented in Table 2.1 show the contrasting text types, sources, and size of text collections used. The majority of studies are shown to have used formal news media from reputable sources such as the Wall Street Journal, Financial Times, Reuters, Bloomberg, Forbes, and Yahoo!Finance [41] [42] [53] [88] [96] [97] [109]. The type of text collected is often general news with only some consideration given to ensuring it is financial or business industry related. Message boards, social media platforms, and company earnings reports have also been used but to a lesser degree than news articles. The number of items varies due to different text types, such as news articles or online messages, and is related to the time frame of the sample and study. The choice of source and text type in the literature are seen to vary in more recent times as data has become more available, more sources have become available online.

The use of content and text analysis to create a proxy for information has resulted in the quantification of vast quantities of news and information. This somewhat new paradigm has meant analysts, policy makers, regulators, and market participants can create a proxy for qualitative information that may contain insights into investor sentiment or the beliefs and intentions of investors regarding investment risk, volatility, and future market movements [9] [21].

The type of information that the finance community has focused on includes company filings, news articles primarily from authoritative sources, and internet messages. The volume of news and timing of news releases are also of great interest to researchers. There exist many examples of studies that look at the timing of news and announcements and their potential impact on financial markets. The timing of the event is as much of interest as the actual content, and in many studies the content of the news shock is not analysed or addressed [5].

Work published in Kelly and Ahmad has shown that the inclusion of a variable that measured and accounted for the timing of a weekly oil inventory announcement improved a baseline autoregressive predictive model for West Text Intermediate crude oil returns [55]. The baseline autoregression model was shown to be improved by including a proxy for news content and the oil inventory announcement.

The content, timing, and information contained in a collection of news or messages are a consideration when investigating the effect of sentiment. Following from the studies presented in Table 2.1, Table 2.2 shows the same studies and the choice of financial data and predictive and explanatory models used to assess the influence of news on financial data. Market indices, particularly the Dow Jones Industrial Average, are used frequently [8] [15] [27] [88] [96]. Firm level data has also been used [2]. A summary of the approaches taken in a number of influential studies that use content and financial modelling show some similarities in the approach of content analysis and financial modelling. Many of the models used to assess the impact of news are regression based while the majority of studies use word level features that are extracted from text using the bag-of-words model.

Schumaker et al. in [88] describe the *AZFinText* system that combines a financial news article prediction system and a sentiment analysis tool to make simple predictions for a trading system. The difficulty of using text based sentiment, or so called "artificial emotions", for prediction is highlighted in Cabrera-Paniagua et al. where an autonomous system is developed to aid decision-making to inform a user about stock market movements [19]. This is again highlighted in [43] where a system is developed to combine various features of the news text, such as word count volume and sentiment analysis, with stock returns to improve stock purchase decisions. A feature selection method is proposed by Nassirtoussi et al. named *Synchronous Targeted Feature-Reduction* that weights semantic and sentiment text features extracted from the headlines of financial news to predict movements in currency pairs in the Forex market [73].

Many of the systems described in this literature review contain varying degrees of integration and automation. One disadvantage to many of the systems is a lack of automatic methods for data processing and modelling. The systems typically assess the impact of news on financial markets through price prediction. With financial and text data becoming more prevalent and freely available through open vendors like Quandl, these systems of analysis are better suited to adapt to this increase in data availability. Systems that can take advantage of these changing datasets and can assess the influence of information extracted from these data in a more autonomous manner will have many novel applications. If it from this observation that motivated the development of a system in this thesis. The system can take advantage of this open data and connect to these varying data sources and combining the analysis of these data with financial data in different quantitative models.

Table 2.2: A number of studies that have combined analysis of financial assets and text analysis methods. The data in each study is at a daily frequency except those notes (*) which include intraday (hourly), intraday (minute and hourly), annual and quarterly, intraday (20 minute), three day window around earnings announcements (**) respectively.

Author	Financial Data	Numerical Method	Period
Wuthrich et al. (1998) [109]	DJIA, Nikkei 225, FT100, Hang Seng, Singapore Straits	k-nearest neighbours, artificial neural networks, Naïve Bayes	06/12/1997-06/03/1998
Peramunetilleke and Wong* (2002) [82]	Exchange rates	Categorical forecast	22/11/1993-27/11/1993
Fung et al. (2003) [41]	Stock prices	support vector machine (SVM)	1/10/2002-30/04/2003
Antweiler and Frank* (2004) [8]	DJIA	Naïve Bayes, SVM	2000
Henry** (2006) [50]	Stock prices	Panel Regression	1998-2002
Das and Chen (2007) [27]	Morgan Stanley High-Tech	Classifier ensemble	07/2001-08/2001
Tetlock (2007) [96]	DJIA	VAR, OLS regression	01/01/1984-17/09/1999
Tetlock et al. (2008) [97]	Stock prices, future cash flows	OLS regression	1980-2004
Li* (2010) [62]	Stock prices, Index and Quarterly earnings, cash flows	Naïve Bayes and Dictionary-Based	1994-2007
Loughran McDonald (2011) [66]	Stock prices	OLS regression, Fama-Macbeth regression	1994-2008
Bollen et al. (2011) [15]	DJIA	Self-organizing fuzzy neural network (SOFNN [60])	28/02/2008-19/12/2008
Engelberg (2012) [35]	Stock prices, short sales	Panel Regression, Fama-Macbeth regression	03/01/2005-06/07/2007
Schumaker et al.* (2012) [88]	S&P500	Support vector regression	26/10/2005-28/11/2005
Jegadeesh and Wu** (2012) [52]	Stock prices	Multivariate regression	01/1995-12/2010
Yu et al. (2013) [110]	Stock prices	Naïve Bayes	01/07/2011-30/11/2011
Jin et al. (2013) [53]	Exchange rate	Linear Regression	1/01/2012-31/12/2012
Garcia (2013) [42]	DJIA	OLS regression	1905-2005
Liu and McConnell** (2013) [65]	Stock prices	Probit Regression, pooled OLS regression	01/01/1990-31/12/2010

This section has reviewed methods of text analysis that have been used in financial literature to generate explanatory variables for financial instruments from news and text data. Many of the studies and systems presented follow a similar structure of analysis and implementation. A market or application area is chosen typically stocks and indices are the most prevalent. A method of extracting information from text is then used, such as dictionary based methods and classification, while a number of different text sources are chosen. Often these choices are subjective and many of the studies do not present an investigation into more than one financial market or consider how different text types and sources will have on computing results. This brings the discussion to the development of a system that can allow different data inputs and uses different models interchangeably. Combining these methods into a system and implementing the necessary functionality to carry out the processing tasks is what has motivated the development of a system that is described in Chapter 3.

2.5 Conclusion

This chapter has presented an overview of literature from a cross section of studies in text analysis, statistical analysis, and econometric methods. The literature reviewed in this chapter has primarily described methods and systems of text analysis and financial analysis and their application to creating a proxy for information. The merits of and details of content analysis methods in finance have been discussed and how these methods have been used with more traditional statistical models of analysis in finance.

Studies in finance that have used content analysis methods have relied frequently on pre-existing systems to do text and content analysis (Table 2.1). Many of these studies have also relied on news from reputable and well known sources also. The information extracted from the various text collections has relied on separate modelling techniques to assess any relationship with financial markets. These studies have relied on regression techniques in some cases to aggregate and model the impact of sentiment and information proxies (Table 2.2). Other modelling techniques that have incorporated sentiment and information derived from news and have been described also. An important aspect to these studies is a reliance on systems to collect, aggregate, processes, and analyse. As many of the studies rely on methods from two separate disciplines a system to bring all the necessary functionality into a single pipeline means assessing the impact of news on financial markets can be estimated more efficiently and reliable.

3 Implementation

3.1 Introduction

In this chapter the implementation and development of a system for generating a time series of sentiment extracted from text and the incorporation of this variable in a statistical model with financial times series data is described. The text analysis and statistical methods implemented and models used draw from the literature presented in Chapter 2. The system contains two main components, the first performs text analysis and the second statistical modelling and follows the general outline:

- Collect, curate, and store unstructured text and financial data
- Apply information extraction methods to generate a time series of affect at different frequencies
- Aggregate and align affect and asset data
- Model affect time series with asset time series
- Determine relationships between affect time series and asset time series using hypothesis testing

The system combines different aspects of content analysis and quantitative modelling with that of data collection and aggregation tasks, ultimately ending in several model estimations. By employing computing paradigms this implementation allows different data sources to be input

into the system facilitating the analysis of different news topics and sources and any financial asset or market. The novelty of the system lies in its ability to examine any news source or text type and combine the output of the content analysis component with any financial time series data. While many studies and systems in the literature subjectively choose specific text and financial data the implementation and methods chosen here allow more data types to be input without close supervision. This allows the influence of different types of news and text in various financial markets to be assessed.

The following chapter will discuss how the methods were implemented with an emphasis on the technologies and resources used. One of the main objectives of the system was to create a work flow that allows two distinct disciplines, financial modelling and text analysis, to work together and secondly automate this process to allow the system to collect, aggregate, analyse, and present the results more efficiently than performing the tasks by hand. In the following section the prototype system is described, first a general overview is given and the two main components of the system the text analysis component and statistical modelling are described in detail.

3.2 System overview

The implementation consists of two main components: a corpus builder and text engine implemented in *Python* and the core statistical engine implemented in *R*. An overview of the system is seen in Figure 3.1.

The prototype system consists of two main components one of which performs text analysis and the other performs statistical modelling using methods drawn from econometrics and time series modelling. Both components are described in more detail in the following sections of this chapter. The text analysis and statistical components of the prototype conduct data harvesting and aggregation tasks. The text engine builds a corpus of text from specified sources, while the statistical component aggregates data from available financial data sources. Data are processed accordingly by each component before being aggregated together by a time ordered index creating the time series data. Any necessary transformations are applied and different model estimations are computed. After the models are estimated the results and output of the approximations are saved to a file and also output to the web GUI component, which is used as a visualisation tool. This system is evaluated in Chapter 4 where two case studies are described and examine the influence of different news types and sources for two different financial markets.

3.3 Text Analysis Component

The components implemented for the text analysis engine contain a number of functions for collecting, storing, aggregating, parsing, and analysing text. The overall architecture of the text analysis component and functions developed is shown in Figure 3.2 and is implemented in

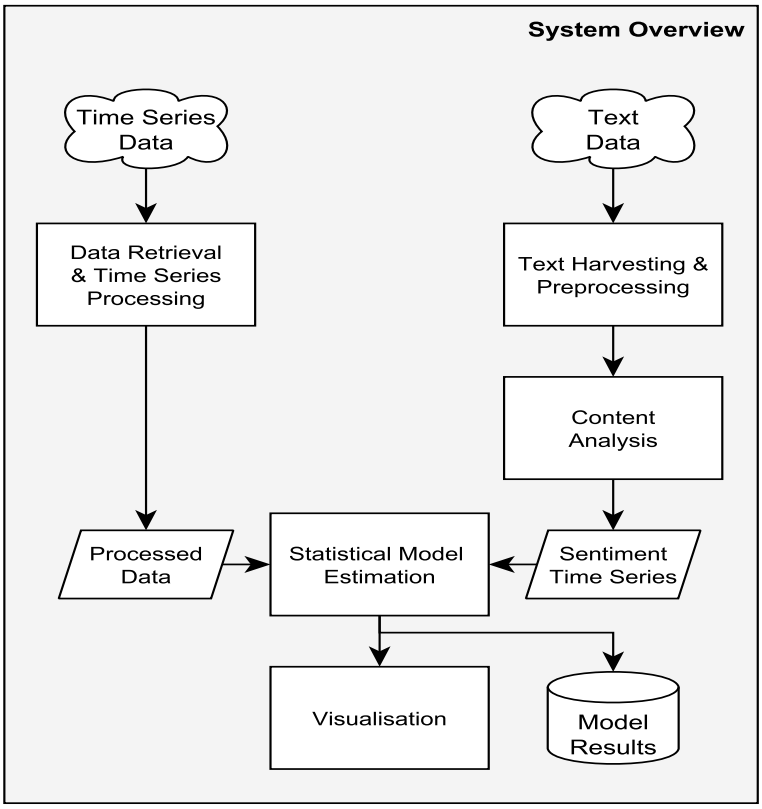

Figure 3.1: An overview of the system implementation consisting of the three main components facilitating text analysis, statistical modelling, and visualisation

Python. The main functionality is divided into two main components, the first component deals with the collecting, structuring, and aggregation of text documents from online news sources for the purpose of building a text corpus. The second component imports this corpus of text and performs content analysis using the bag-of-words model and a given dictionary of categorised terms. The output of this component is a time series of sentiment extracted from the corpus of text and is organised in time. This time series is then passed to the second part of the system which aggregates the time series output with financial data and performs the statistical modelling. The individual functions of the text analysis component are described in more detail in the following section (Section 3.3).

3.3.1 Text Preprocessing and Data Scraping

Collecting structured and unstructured digital text using a computational process means a large volume of text can be collected relatively easily. Using any accompanying metadata and structuring the text in a consistent manner, such as through a schema like XML or JSON, a corpus of text can be built and processed more efficiently than if left unstructured. If the time

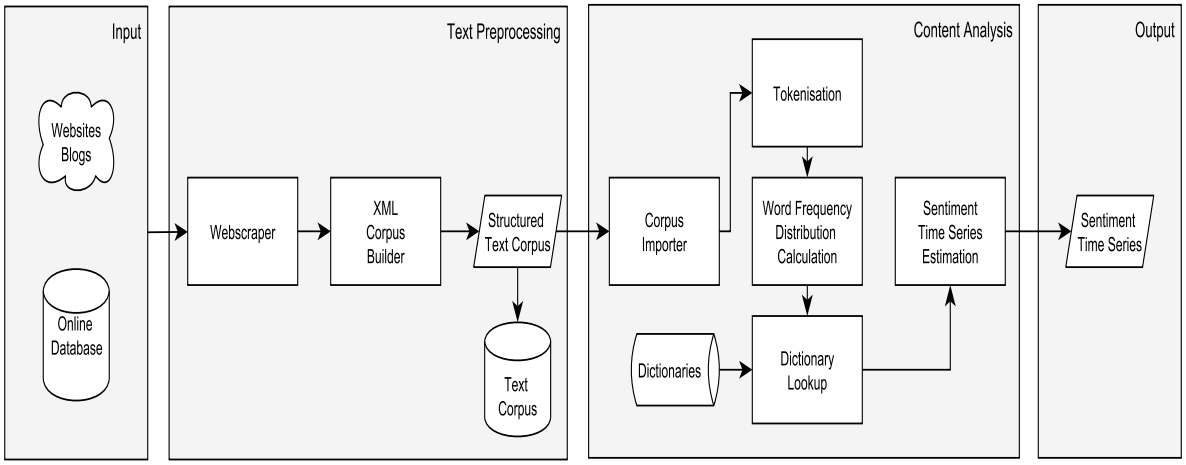

Figure 3.2: A system diagram of the text analysis component developed for the system that computes a time series of sentiment

of publication is stored in the metadata then the text can be filtered and stored chronologically. This is the essential starting point for the content analysis system and method implemented here for developing a time series variable. This time series variable computed by the system acts as a proxy for sentiment contained in the text. This functionality is illustrated in the text preprocessing section of the architecture diagram in Figure 3.2.

The system is capable of collecting text from different online sources including websites and blogs (Financial Times, Oildrum blog) and online databases and archives (Proquest, Lexis Nexis). Text from these sources are formatted and structured in the same way using XML to store metadata for each text document. Each text document contains XML tags to mark the date, source, title and body of text for each individual news article. An example of this scheme being used to structure a text document is shown in Figure 3.3. Each article collected was tagged according to the illustrated XML scheme and organised chronologically with other articles to form the text corpus.

Webscrapers were written to navigate and download news articles from the Financial Times (FT) and Oildrum blog websites. Each webscraper was written in Python and used the *BeautifulSoup* screen-scraping library, which contains an HTML and XML parser and builds on top of the *urllib2* Python library for opening URLs and handling HTTP requests. The Oildrum blog has an archived news section where each article published on the blog was stored according to the date of publication. A URL to the archive page is given and allows the web scraper to connect to the site. Additional functions were written to allow the scraper to navigate the archive downloading news articles and storing the associated metadata. Each article collected was formatted in XML with the metadata describing the article title, text, time of publication, and source.

```
<?xml version="1.0"?>
- <article>
    <source>Financial Times</source>
    <article-datetime>21 May 2013 12:30:00 +0000</article-datetime>
    <title>ICBC / Goldman Sachs: farewell</title>
    <text>What does it say when Goldman Sachs no longer wants you but others
      rush in? The US bank has sold its remaining stake in ICBC, the worlds
      biggest bank by market capitalisation, to realise almost three times the
      initial investment. In truth the deal does not say much more than thats
      private equity for you. Goldman Sachs paid $2.6bn in 2006. In five sales,
      including this weeks, it has netted profit of $7.3bn for a compound annual
      return of 25 per cent. Nice. Since Goldman took a pre-listing investment,
      ICBCs assets and revenues have nearly trebled, while profits rose by a
      factor of six. Since it listed in October 2006 its shares have gained two-
      thirds. But it has not all been smooth sailing. The volatility of the stake,
      marked to market prices, provided a roughly $500m drag on two occasions,
      although on average it added that much each year. And the profit does not
      account for the costs of the Goldman expertise that has been shared with
      ICBC. Nor does the bank take all of the proceeds: some of the stake was
      held for clients. It is time for Goldman to move on. The genesis of the stake
      came from Goldman's private equity team and now ICBCs growth is slowing.
      The banks price, relative to book value, has slid from more than two times
      on listing, and a peak above 4 in 2008, to 1.2. Revenues are expected to
      grow about 10 per cent a year over the next two decent, but half the rate
      enjoyed during Goldmans investment while profit growth of 6 per cent is a
      fifth the rate the US bank enjoyed. On a private equity basis, the deal
      performed. It should be remembered that when former Goldman boss Hank
      Paulson first cosied up to ICBC in 2003 by buying portfolios of bad loans,
      ICBC had an official non-performing loan ratio of 25 per cent. Its success
      was not certain. Goldman took a risk and got rewarded. That is just how
      private equity should work. Email the Lex team in confidence at lex@ft.com
      </text>
</article>
```

Figure 3.3: An example of a news article from the Financial Times that has been structured using XML and the accompanying metadata

Another scraper was written to navigate the FT website to collect articles based on a given key word query and find the most up to date news published about a particular topic based on the query. The supplied keyword query can be used by the scraper to access the FT's search function of the website's archived news articles. The web scraper can then navigate the returned results collecting the news articles. The web scraper built for the FT allows news articles about a topic to be collected as soon as they are released, the scraper can be left to run continuously and check for newly released news articles. Due to licensing, permission may need to be sought through a service such as the Financial Times Developer Programme in order to allow this real-time data collection and scraping. For the online databases Proquest and Lexis Nexis, news articles can be downloaded in bulk by specifying a query, access to these databases was covered under an academic license. This process can also be automated using the webscrapers developed with the BeautifulSoup library and the screen scraping library *Scrapy*, also implemented in Python. More recently, both companies have opened access to their services using developer APIs but require special license agreements to use these services, as a result of this restriction an implementation to their APIs has not been implemented to date.

Once the text has been collected and structured using the XML scheme, it can then be stored in a time organised corpus and backed up in a database. Storing the text and structuring the corpus in this way allows articles to be quickly aggregated according to time and analysis to be performed more easily. Articles are organised in a diachronic manner in the corpus as much of their use is in the construction of a time series. The corpus of text can then be loaded into the content analysis component of the system.

3.3.2 Content Analysis for Domain Text

Content analysis is used in the text analysis component of the system to extract and measure information from the text corpus. The frequency of pre-categorised terms (negative or positive terms for instance), defined in the supplied dictionary, in a text can give an indication to the tone of the text and underlying sentiment. In some cases, in domain text for instance, the meaning of a term may change depending on the context. A term that is typically negative in a general language context may have another meaning in a domain text. To investigate this potential misinterpretation of terminology, a general language dictionary (the General Inquirer dictionary) is used in conjunction with a domain glossary (domain-specific lexica), how this is incorporated into the content analysis component is described in this section. An algorithm is presented in the following section that summarises this process and has been implemented in the system (Algorithm 1). The algorithm adapts the standard bag-of-words model of content analysis to account for domain words and phrases.

The content analysis component implemented in Python (Figure 3.2) follows a typical text processing pipeline and uses methods from the Natural Language Tool Kit (*NLTK*) API [10] to perform text processing tasks. The resources and tools in the NLTK API allow efficient text processing and analysis to be performed and is a fully open sourced API that can be incorporated or modified for use with other systems. From Figure 3.2 the first function of the content analysis method loads the text corpus and parses each news document using the inbuilt XML parser in Python. The text of the title and body of the news article are read and stored according to the date of publication that is specified in the *article-datetime* tag, which is the time of publication for the news article.

This text is then passed to the text processing function, which performs tokenisation using the *RegexpTokenizer* from the NLTK library and splits the strings into sub-strings using a regular expression. Tokens can be separated based on whitespace, line breaks, or by punctuation characters and other rules can be specified. Tokenisation allows the text to be split into tokens based on a separator, in this case whitespace is used, and returns a vector of terms for the document processed. Text normalisation is also performed which removes non-necessary characters. The removal of non-alphanumeric characters, additional white space, and symbols may be among the less complex tasks of normalisation. Normalisation that is performed here includes the removal of non-text characters and non-standard punctuation and setting the re-

sulting tokens to lower case. Additionally, text can be broken into different $n-gram$ sequences at this stage of the processing. An n-gram sequence that is extracted from text can consist of elements such as syllables, characters, letters, or words. A single n-gram is often referred to as a unigram, with an n-gram of size two being a bigram.

The option to perform further pre-processing steps such as stemming is also available through functions implemented in the system. Stemming will reduce the inflections of words to their stem or root form but not necessarily the true or morphological root. It is possible then to investigate the effect that different word senses may have on the count of terms. A stemmer may be used on the raw text after tokenisation. Often the stem is not the actual root word but a stem which related words or derivatives can map to. A typical stemming algorithm will incorporate a set of stemming rules [78] [83]. The Porter stemming algorithm [83] also allows the removal of common morphological and inflexion endings from tokens and is used briefly in the evaluation of the system. The *PorterStemmer* function from the NLTK library is used and wrapped with helper functions in the system to allow stemming of text. Functions to perform part-of-speech tagging and removal of stopwords are also present but not used by default or evaluated here unless stated otherwise.

The frequency distribution of the words in the text can then be calculated by passing the tokenised document (word vector) to the next function. This has been implemented using the *FreqDist* class from the NLTK. This creates a Python dictionary object containing each word that has occurred and it's frequency of occurrence in the tokenised document passed to the function. The frequency distribution of multiple words or sequences (n-gram sequences) occurring together (collocations) can also be computed using the *bigram* or *trigram* functions based on the *FreqDist* class or extended to a custom number of n-grams.

The frequency distribution of words is then passed to a dictionary lookup function that reads in a dictionary of categorised terms, the General Inquirer dictionary has been integrated as a default. This component can also incorporate more dictionaries or a glossary of domain terminology. The dictionary lookup function can query the word frequency distribution for the occurrence of terms from the dictionary. The count of each term can be aggregated with all terms in a category giving an overall frequency occurrence of a category in a text document. This count is then normalised by the total number of words occurring for that time period (in the case of a time series) or in the text document. This gives the relative frequency and is useful in that the score is less sensitive to unusual changes in news volume. The relative frequency can be defined by saying that for k number of w words in a category of a dictionary, the total frequency of their occurrence in a corpus of text D can be denoted as the summation of $f(w_D^k)$. If we denote time as t, then the frequency of all words in a category of a dictionary that occur in all articles published during some time period (during one day for example) can be written as the sum of $f(w_t^k)$. If the frequency of a word or token n that occurs in all articles published during a single time period is the sum of $f(w_t^n)$, then the relative frequency of a category of terms from a dictionary is calculated as:

$$\text{Relative frequency} = \frac{\sum_{i=1}^{k} f(w_t^k)}{\sum_{i=1}^{n} f(w_t^n)} \tag{3.1}$$

where:

n = the total number of tokens in all documents published during a time period
k = the number of terms in the dictionary category
$f()$ = the frequency or count of a token
t = the time period

The category that is predominately used and evaluated from the GI dictionary is the negative category. This category contains a unique list of 2,005 single word terms. It is the main sentiment or affect category studied in the domain of finance, primarily due to theories of overreaction from investors to negative news, the so-called asymmetric response to news.

Previous authors have cautioned that a misclassification of terms may occur when using a general language word list, such as that used in the General Inquirer, in a domain specific text such as finance and business news. Loughran and McDonald produced a financial negative and positive word list that they determined to be more accurate in identifying polarity in financial text [66].

The system is capable of accounting for domain words and phrases by incorporating a domain glossary with another base dictionary such as the GI dictionary, which is used by default. This can be beneficial when analysing domain text. A term such as "crude" categorised as negative in the GI word list, would be frequently used in oil related news, but usually to refer to "crude oil" and not interpreted as negative as in general language text. Terms can represent the knowledge of a domain and a glossary of domain terminology can be used to add knowledge when examining topic specific news or text. A domain glossary can then update a general language list of terms, such as the GI categories, and help refine term labelling.

To evaluate the system's content analysis method for domain text, data from the commodities markets were chosen (Section 4.3). The system incorporates two domain glossaries for the oil industry these include the Platts[1] glossary, and the Oil and Gas UK glossary[2]. These lists cover common terms and abbreviations from the oil and energy industries. Both sources are seen as being reputable; Platts is a leading source of information on energy markets and provider of benchmark prices, while Oil and Gas UK is the leading trade association for the UK's oil and gas industry. The Platts glossary contains 704 words with expanded abbreviations, while the Oil and Gas UK glossary contains 133 words including expanded abbreviations. The count of affect terms will not be altered if GI negative and positive terms occur independently of the domain glossaries. If the term "crude" occurs separately from the compound term "crude oil", the latter will not increase the negative word count but the former will still be accounted for.

[1] wwww.platts.com/glossary
[2] http://www.oilandgasuk.co.uk/glossary.cfm

This method of disambiguation for specialist text has been incorporated into the dictionary lookup function (Figure 3.2) of the content analysis component. This procedure is summarised by Algorithm 1 and is implemented in the system so as to allow different dictionaries to be imported and used together.

Algorithm 1 Generate a time series of affect for a domain specific corpus

Input: Base dictionary C_m consisting of m number of C categories
Domain Dictionary C_s consisting of s number of C categories
Corpus of text documents D organised into n time periods

Output: Time series $F_C = \{f_1, ..., f_n\}$ where $f(n)$ is the sum of the word frequencies during time period n for category C

for $i = 1$ to n **do**
  $w_i \leftarrow$ Tokenise(D_i)
  for all C **in** C_m **and** C_s **do**
    if w_i **in** C_m **and** C_s **then**
      $f(w_n^s) \leftarrow$ Count(w_i)
      $f_n^s \leftarrow \sum_{j=1}^{i} f(w_i^s)$
    end if
    if w_i **in** C_m **and not** C_s **then**
      $f(w_i^m) \leftarrow$ Count(w_j)
      $f_i^m \leftarrow \sum_{j=1}^{i} f(w_i^m)$
    end if
    if w_i **in** C_s **and not** C_m **then**
      $f(w_i^s) \leftarrow$ Count(w_i)
      $f_i^s \leftarrow \sum_{j=1}^{i} f(w_i^s)$
    end if
    $f_i \leftarrow \{f_i^s, f_i^m\}$
  end for
  $F_C \leftarrow \{f_1, ..., f_n\}$
end for

In Algorithm 1 content analysis is performed on all documents that occur during a given time period. In testing and evaluation in Chapter 4 the daily frequency is chosen as financial data at this time frequency is freely available. Daily news articles can be more easily aggregated to this level to give a more consistent time series with less likelihood of missing observations. The input to Algorithm 1 consists of a dictionary and corpus of text. A general language dictionary is used, in the evaluation of the system and by default the GI dictionary is used, and becomes the *base dictionary* consisting of several categories of terms. A second dictionary or glossary of terms can be input also and used in conjunction with the base dictionary. This domain dictionary can account for potential misinterpretations of the base dictionary terminology for domain text.

The corpus of text, another input, is organised in time. This is done by aggregating news and text published for a particular day for instance (for daily time frequency) and organising the aggregated articles by date. The dictionaries and text corpus are passed to the algorithm and for each day (for daily time frequency, any time frequency can be used) all news articles with a time stamp for that day are aggregated together and tokenised into n-gram length words (Algorithm 1 line 1,2). For each category in the supplied dictionaries the frequency of occurrence of terms in each category in the corpus is calculated (line 3 - 16). If a word from the corpus appears in both the base and specialist dictionary, then the specialist dictionary takes precedence and the count for the category in the specialist dictionary is increased (line 4 - 6). The frequency of occurrence for all the specialist dictionary words that occur in all articles that day is computed and added to the overall frequency for the corresponding category in the specialist dictionary (line 6). This function is used again if the word from a text appears exclusively in the specialist dictionary (line 12 - 14). If the token from the text only appears in the base dictionary then the frequency count is increased for this corresponding base dictionary category (line 8 - 10). The time series of term frequency for the category is computed and is the final total frequency score for that day (line 16). The daily frequency of a category is then aggregated into a time series and output by the algorithm (line 18).

When analysing general finance news, such as the *Abreast of the Market* column from the Wall Street Journal, just the GI dictionary is used to estimate negative sentiment from the text. While examining the oil related text corpora, both the GI dictionary and a glossary of oil words are used to summarise the negative sentiment in text taking into account potential miscounts and misinterpretations of domain terms and phrases. The frequencies of the terms from the dictionary category in the corpus are aggregated over a frequency of time. This frequency is for a day in the evaluation and use cases presented in Chapter 4. The overall frequency of the dictionary category appearing in all articles in the time frame (during the day) can then be divided by the total number of words that appeared in all articles published that day to get a count of the relative frequency. Normalising the absolute frequency count of the dictionary category in this way allows the volume of news to be taken into account so large increases in the volume of news that might occur may not impact the frequency score as much. The consistency of each text corpus is also typically checked to ensure that there are not a large number of days with no news. The final output is a time series of sentiment extracted from the text corpus and organised by time and can be passed on to the statistical analysis component of the system to estimate any inter-relationships with financial market data.

The resulting time series can be output in comma separated format and imported into an aggregator function in the statistical analysis component that is written in *R*. All software and libraries used are open source and have free software licenses. The Rocksteady Affect analysis system developed at Trinity College Dublin was used initially to confirm the results of the case studies presented in this thesis and to compare and evaluate the output of the implementation of the text analysis system described in this section. The output of the content analysis component

was found to be consistent across the two systems.

3.4 Statistical Analysis

After the specification and collection of data, exploring the individual data series is the starting point of exploratory data analysis. A univariate analysis of a series is typically carried out before a multivariate analysis is conducted. Exploring the stylised facts of a series can help verify the data and aid in constructing more complex models that provide useful information. For multivariate analysis, vector autoregression (VAR) is chosen as the model to determine any inter-relationships between variables in conjunction with rolling regression and hypothesis testing. The VAR model computed by the system is influenced by the literature and described in more detail in Chapter 2. The VAR model includes control variables that attempt to account for known market behaviour and anomalies such as volatility for instance. The system computes the VAR model giving an indication of the average impact of sentiment. The attributes of financial series are known to change in time and as such rolling methods are incorporated into the modelling component to assess how market changes can influence the results found. Hypothesis testing is also used to examine the statistical and explanatory information of the sentiment variable produced. The modelling component computes different statistical estimations in order to determine the statistical power and quantitative impact that news has on financial markets. In the following section the construction of the statistical analysis component is presented and how the system merges the output of the text analysis component with financial time series in order to assess the impact of news on financial markets.

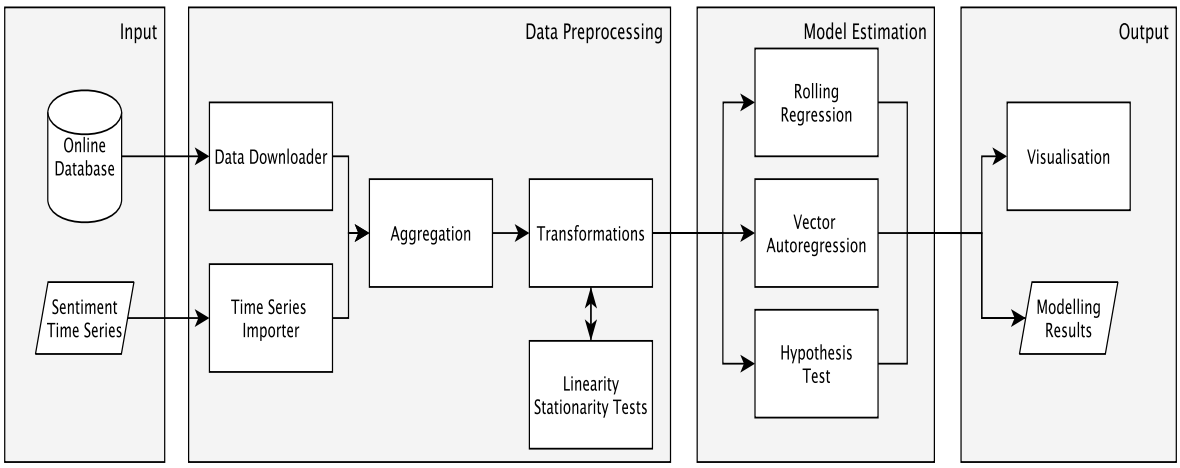

Figure 3.4: A system diagram of the Statistical processing component developed which performs data processing and modelling

An overview of the statistical analysis component is shown in Figure 3.4. The functions and

models in this component were implemented in *R*. The main functions are divided into a data processing component and a modelling component. The data processing component performs the retrieval of financial data and aligns this with the imported results from the text analysis component. The modelling component runs different models to assess the impact that the sentiment variable will have on changes in the price of financial assets. These data and output are passed to a web based visualisation, the full output from the models is also saved to a file.

3.4.1 Data Processing

The first function of the data processing component retrieves financial data by wrapping the *Quandl* API to connect with their database and download queried data. The *Quandl* database collects and stores financial and economic data and provides an API for most major software languages. The data processing component also allows other time series data to be imported (such as data output from the Tradestation Trading Platform) with headers for each column and organised by date. The *Zoo* time series package in *R* is used to process each of the time series and is also used in importing the sentiment time series from the text analysis component. This package is also used for several functions that merge and align the time series data. The data scraping component can be specified to select financial data using a given ticker symbol. Once a data call has been made to Quandl or another database the data downloaded into the *R* environment is ordered and checked if a large number of missing values are present. The time series data are then passed to the aggregation component of the implementation. In this function different time series are collected, aggregated, and aligned according to date. All time series data are aligned to the financial time series (DJIA, WTI), this means weekends and public holidays are dropped as no trading happens on these days. Dates where data does not exist for any of the time series variables in the aggregated data matrix are dropped to ensure the best possible alignment of available data. Care is given to ensure the highest quality data are used from reputable sources and output from the text analysis to ensure sufficient news is collected to generate a time series without a large number of missing days. Daily data are used typically and in the evaluation of the system in Chapter 4 as this time frequency allows the best alignment of financial and news data.

All necessary transformations are performed using functions built with the *R* base library. These transformations include calculating returns for financial price data, z-score values for sentiment, aggregating data to other (lower) time frequencies if required, or other transformations that may be necessary to create a stationary time series. All transformations used by the system for different models are also described in the evaluation of the system in Chapter 4.

The first function of the transformations component calculates financial returns from the price series of the downloaded financial data. The financial data downloaded will include several columns of time series data such as trading volume, or high and low price values for a time period. For a single observation the price quoted from an exchange is often the final price or

transaction price for that time period. For a daily price series, the price quoted for a particular day will be the price at the end of that trading day. This close price is what is extracted from the financial data and aggregated with the sentiment data. There are several methods of calculating the returns from prices. Returns are used in statistical modelling for a number of reasons including normalisation; where asset prices of unequal value can be transformed into a metric that is comparable. There exist statistical properties of returns that make them more amenable in statistical modelling such as being more numerically stable (constant mean, removal of trends). The function takes the closing price time series and calculates returns as follows:

$$r_t = ln(\frac{P_t}{P_{t-1}}) \tag{3.2}$$

For a period t, the price of a financial instrument or index value at this time is P_t. This period t can consist of different time frequencies depending on what series is supplied such as monthly, daily or intra-daily. Using a log of a series with differencing, a series will become a percentage change between each period, obtaining the log will turn the absolute differences into relative differences.

Another function in the transformations component calculates the z-score of a given time series. This transformation is performed on the sentiment time series after aggregation with the financial data. When using variables with different units of measurement, it is often necessary to standardise the series in a similar way to help interpretation. For the sentiment time series, which is comprised of the relative count of negative terms, it can be difficult to interpret any relationships of this series with financial data. By calculating the z-score, the series is standardised to have zero mean and unit variance, the distribution of the series is centred on zero. Any results are thus interpreted as the change relative to the mean, results from a regression would be interpreted as a unit change in the standard deviation, a function in the system calculates the z-score for a time series using Equation 3.3.

$$z = \frac{x - \mu}{\sigma} \tag{3.3}$$

where:

x = the time series or vector of data
μ = the unconditional average of x
σ = the unconditional standard deviation

Model estimation using z-scores is beneficial when comparing values of different units. It is however important to ensure the mean and variance do not vary considerably over the entire sample set as the resulting z-score interpretation may be effected as a result. All sentiment series used in the case studies and evaluation presented in Chapter 4 are transformed and interpreted as z-scores.

The last component of the data pipeline can work in conjunction with the transformations component, this component tests whether the time series are linear and stationary. Non-stationary data will often have trends or cycles occurring in the data which results in changes over time in central moment behaviour, changing mean, variance, and covariances. Due the presence of these patterns, it is often unpredictable and difficult to model these data with linear models. An awareness of the anomalies of the series will help in determining what transformations may be necessary to obtain a stationary series. For instance, by de-trending a series, a deterministic trend can be removed by simple subtraction that does not affect the observations of the series but instead may transform the process from non-stationary to a stationary one. This allows the transformations component to be informed by the stationarity tests.

To obtain a stationary series it is necessary to have time-invariant first and second moments, having constant mean and variance. Strict stationarity requires that the variable distribution be invariant in time. Weak stationarity assumes that the mean of r_t and covariance between r_t and r_{t-l} is time invariant, where l is an arbitrary integer. The augmented Dickey-Fuller test (ADF) is used to see if a time series has a unit root, with a negative value indicating the degree of rejection of the hypothesis that there is a unit root. This test is implemented using the *adf.test* function from the *tseries* package in *R*.

3.4.2 Model Estimation

After data processing the time series data are passed to the model estimation component (Figure 3.4). In this component a vector autoregression, rolling regression, and hypothesis test are computed. The results and output of the models are stored and can be passed to an interactive GUI front end where some of the results are displayed. This is the final component of the system.

The processed time series are passed to a vector autoregression (VAR) model that estimates the model and regression coefficients for each of the included variables. The details of this model have been described previously in Section 2.3.3. The regression equation to be estimated can be specified prior to running the system. The system can update the data series and re-estimate and update the models periodically. In the VAR model estimation, a prior equation is supplied that is used in the evaluation of the system (Chapter 4) and has been drawn from the literature. These models and equations are described in more detail in the case studies in Chapter 4. These models rely on financial data, a sentiment time series, and additional control variables. The processing of financial data and construction of the sentiment time series has been discussed in the previous section. The model estimation component contains functions to generate the control variables that are included in the model, an example of these controls include a Monday and January dummy variable. A function in the model estimation component takes in the date index of the processed data and creates a dummy series for Monday and the month of January to be used in the model estimations.

The modelling component estimates linear autoregression models and lags of the dependent

and independent variables need to be included in the regression estimations. The order of the variables can be examined closely by using the autocorrelation functions and an information criteria such as AIC both of which are described in more detail in Section 2.3.2. The auto and partial correlation functions can be estimated using the base functions in *R*. Typically for the models estimated by the system five lags of the variables are included in the regression equation. This is to account for up to a week of trading data as specified in the literature and has motivated the evaluation of the system in Chapter 4.

An informal method of testing the changing impact of news can be performed using a rolling regression. An ordinary least squares, an hypothesis test, and a moving window function to sample input data at different times are the basis of the rolling regression model computed by the system. It is a simple solution as it does not impose any particular assumptions or structure on the changing attributes of the series. For a window of width n, a rolling linear regression model can be expressed as:

$$y_t^T = \phi_0^T + \phi_1^T x_t^T + \epsilon_t^T \tag{3.4}$$

where:

y_t = a single vector of observations
x_t = an explanatory variable
T = the number of windowed periods for the full sample set, where n is the window length

Rolling window methods can be used to assess how consistent a model's estimation is in time. If parameters change over the sample set, the rolling estimates will capture this instability. The modelling component will compute the regression model for each period n and also computes the hypothesis test for the significance for the inclusion of sentiment in the model. The hypothesis test is described in more detail in the subsequent section (Section 3.4.3). The length of n chosen is for 250 observations (250 days for daily data), which equates to a typical year of financial trading data. The output from the model includes a time series that includes the regression coefficient values and their accompanying p-values and a time series of binary values indicating the significance of the hypothesis test. The result of the hypothesis test is a time series graph that is output and shown in the evaluation of the system in Chapter 4.

3.4.3 Model Diagnostics

After the data have been processed and passed to the model estimation component the regression coefficients are computed along with the statistical significance value for each coefficient. Three parameters are output with each model estimation that describe how well fit the model is, the adjusted *R*-squared value that indicates the amount of variance explained by the model ($\bar{R}^2$), the result of the hypothesis test for the sentiment variable and whether it has explanatory information (X^2), and AIC values for each model so comparisons can be made between similar

model fits. The AIC value is used as a comparison between similar models and shows if the addition of an extra independent variable helps the model specification. If the value is less after the inclusion of the independent variable relative to the initial model then it is considered a valid variable to include. The statistical significance (measure by the P-value) of this additional variable in the model can also support this conjecture.

The AIC is not indicative of much meaning by itself but is instead used when comparing one model to another, when two or more models attempt to explain the same data or variable. The addition of new variables and their explanatory power for a model can be seen in the AIC measure. Although it is difficult to compare the AIC value between distinctly different models, similar models can be compared using the AIC to determine the overall improvement of additional components or individual variables. A lower AIC value is often favoured when comparing multiple similar models. When considering the AIC value for a model i of a set of models, where the model with the minimum AIC value is $minAIC$ the AIC values for all models in the set can be rescaled accordingly by computing $\Delta_i = AIC_i - minAIC$. This value is computed by many standard statistical libraries that compute regression models and the parameter can be extracted and stored by the model estimation component.

The R^2 value is also computed to summarise the goodness of fit of a model. The adjusted r-squared is the ratio of the residual sum of squares (difference between observed and predicted or fitted values for the model) versus the total sum of squares (difference between the observations and mean of the data). The results output from the model estimations in the system component look at the adjusted r-squared value $\bar{R}^2$, which is the ratio of the variance of residuals to the variance of the data sample. The $\bar{R}^2$ for a model is defined as:

$$\bar{R}^2 = 1 - \frac{VAR_{res}}{VAR_{total}} \tag{3.5}$$

where:

VAR_{res} = the variance of the residuals of the sample
VAR_{total} = the total variance of the dependent variable

This statistic can adjust for the number of variables being included in the model and is computed when performing the regression estimation. The modelling component will also extract and store this value from the results of the regression model.

Hypothesis testing is used in the system to determine the statistical significance of including the sentiment variable to the model to explain financial returns. The hypothesis test is used to examine whether one variable and its accompanying lags can be used to explain variation in another variable. This can be done by testing the statistical significance for a group of coefficients. The system evaluates the relationship between news sentiment and financial returns by testing the hypothesis that sentiment has statistical power in explaining variation in returns. The null hypothesis states that all coefficients of lagged variable of sentiment are equal to zero.

The F-statistic is calculated and compared with a significance level or critical value. This significance level is often chosen to be equal to 0.01, 0.05, or 0.1. The chi-square statistics (X^2) can provide a quantitative measure of this relationship. This statistic measures the difference between the expected and observed values between variables.

The F-statistic is the ratio between the explained variability and is calculated as:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} \tag{3.6}$$

where:

RSS = the residual sum-of-squares fit for unrestricted model (1) and restricted (0)
p_1 = number of parameters of the unrestricted model
p_0 = number of parameters of the restricted model

The F-statistic measures the change in the residual sum-of-squares for an additional parameter or group of parameters in the unrestricted (larger) model versus the restricted (smaller) model. If the F-statistic is found to be greater than the critical value, then the null is rejected, the larger the value of the F-statistic than the more useful the variable under question is. For the modelling component this is tested on the sentiment variable and its inclusion in the model. The hypothesis test used by the system to evaluate the impact and significance of sentiment is defined as:

$$H_0(Null) : \beta_1 = \beta_2 = \ldots = \beta_n = 0 \quad \text{Sentiment does not impact returns}$$

$$H_1 \text{ (Alternative): } \beta_1 = \beta_2 = \ldots = \beta_n \neq 0 \quad \text{Sentiment does impact returns} \tag{3.7}$$

where:

$$\beta_n = \text{the coefficient value(s) estimated for the sentiment variable}$$

The test implemented in the system presented uses the *linearHypothesis* function from the *R* package *CAR* (companion to applied regression). This function can carry out a Wald-test based comparison computing the F-statistic or asymptotic chi-squared statistic between the estimated regression model (the unrestricted model with sentiment) and a linearly restricted model without the sentiment variable. The chi-squared statistic (X^2) has been used in Tetlock [96] and is used by the system here to allow a comparison to the results in the literature.

Traditional OLS regression is based on the assumption that the series being modelled is stationary with no heteroskedasticity present. If heteroskedasticity is present, the coefficients of the regression will not be biased, but the estimate of the variance of the coefficients will be biased. The standard error in this case, calculated from the sample standard deviation of the dataset, will be a biased standard error and alter the statistical significance and any inference made from the regression coefficients computed. Heteroskedastic-consistent standard errors can be used to adjust the standard errors correcting for heteroskedasticity without affecting the

coefficient value. Heteroskedasticity and autocorrelation can be somewhat accounted for using adjusted standard errors and is frequently used in econometric modelling and when dealing with financial returns [42] [96]. Adjusted errors are used for all the regression estimations in the modelling component of the system. All regression results presented use Newey-West standard errors [74] to give a heteroskedasticity corrected covariance matrix in the model estimation. The *lmtest* package is a diagnostic package used for regression modelling and is used in the modelling component of the system to adjust the standard errors along with the *sandwich* package for econometric computing, which produces an adjusted estimate of the covariance matrix of the parameters of the regression model.

The output and results of the model component and data from the data processing component are saved to a file and passed to the final component of the system which creates an interactive visualisation of the results and is discussed in the next section.

3.4.4 Visualisation

The visualisation component is the final part of the system and consists of a GUI interface built in *R* using the *Shiny* package and allows a clean interface to be wrapped around the statistical analysis component of the prototype system. The GUI component is based on a JavaScript and HTML framework and allows the creation of a user interface with server side capabilities. All applications built using this framework use web based client side and server side design principle. Using the *Shiny* package a user interface script must be created that controls the layout and appearance of the application (client side). A *server* script must also be created to act as the controller of the application (server side). A final *helpers* script is also used with the server script to provide additional logic, data retrieval, and functions that are necessary for data and results preparation for the GUI. Much of the framework acts as a wrapper for *R* functions and scripts, where more detailed functionality can be specified and then mapped to the UI created. The goal of the *Shiny* framework is to allow users to showcase data analytic work with a minimum overhead of GUI design and deployment of a web application.

The web based framework makes deploying the system to a web server with a front-end possible and allowing the prototype to be more portable and work independently from a user or individual computer. Figure 3.5 shows the front-end that can be accessed from many popular browsers and hosted on *RStudio* server products. No native mobile support exists for the *Shiny* framework currently, although by hosting the application on a non RStudio/ Shiny server an application can be viewed through a mobile web browser. The panel of the left allows a user to input a ticker symbol and retrieve financial data from the *Quandl* database. The panel on the right displays the retrieved financial data and also the sentiment time series computed by the system.

The interface seen in Figure 3.5 allows a user to change and query the data at different time scales and to visualise the sentiment time series generated by the text analysis component.

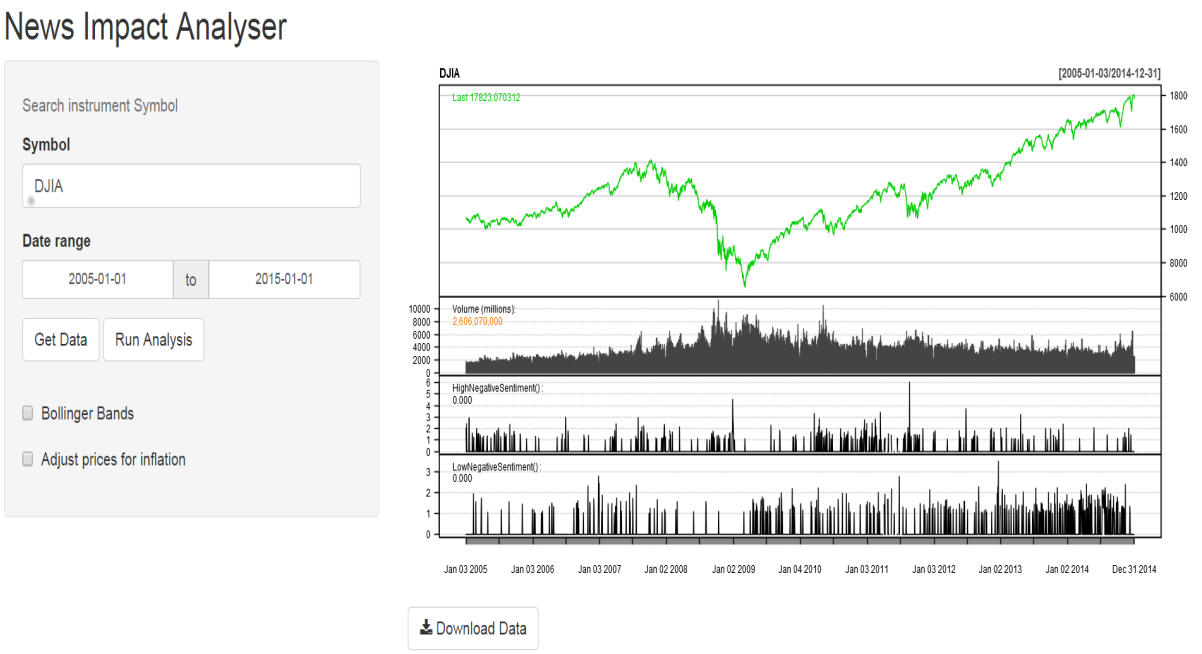

Figure 3.5: Front-End of Prototype showing data retrieved from Quandl and the level of negative sentiment calculated from the FT *Lex column*

The visualisation relies on the *Quantmod* package in *R* to display the time series graphs using the *chartSeries* function for creating financial charts. The panel on the left allows a user to change the time frequency and add technical indicators, the results of which are then graphed on the time series plot. A custom indicator was created to process and display the results of the sentiment time series and was added to the front-end using the *addTA* function in Quantmod. This was done by creating two functions in the helpers script that take in the output from the text analysis component. To create the displayed sentiment time series, the z-score transformed sentiment time series was used and the absolute value of the sentiment series was computed. Observations from this series that were one standard deviation above (*HighNegativeSentiment()* function) and below (*LowNegativeSentiment()* function) the average sentiment for the sample period made up the high and low sentiment series seen in Figure 3.5. These time series charts can give an indication to the level of sentiment (negative sentiment in the example shown) in a news stream.

An analysis can also be run that computes a VAR model with the returns of the financial time series specified from the front end and the sentiment time series computed. The VAR model estimated by the system by default is defined and described in Section 4.2.2. The resulting regression coefficients that summarise the impact of negative sentiment on the financial returns and their statistical significance are tabulated and displayed as seen in Figure 3.6. From the figure the lagged negative sentiment variables are shown with the coefficient value estimated by

the model in basis points and the p-value for each of the coefficients.

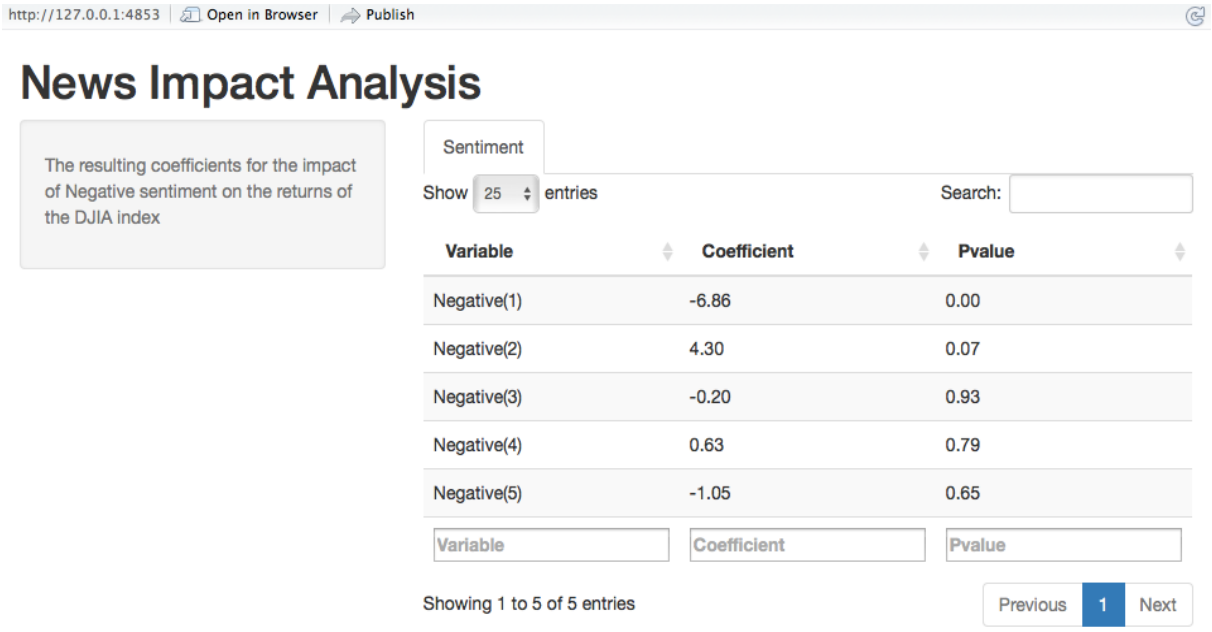

Figure 3.6: Resulting coefficients from robust regression analysis of negative sentiment extracted from the FT's *Lex column* impacting DJIA returns

One of the main differences in this implementation and previous studies that have used content analysis with statistical modelling is in the automation of the processes needed to perform the final analysis. Where previous studies have had to collect and process text and financial data by hand, the system described in this chapter removes this burden from the user. By modularising much of the processes and creating functions to perform this work, the error from performing these tasks manually can be reduced. The automation of the data processing components and modularisation of these functions allows new inputs, both text and financial data, to quickly be added to the system facilitating an easier investigation into the relationship that different news topics and types can have on different financial assets in a number of distinct markets.

3.5 Conclusion

This chapter has described a system for generating a sentiment time series variable that can act as a proxy for information contained in text and news. The system uses this sentiment variable in a statistical model to determine any potential relationships with changes in the price of financial assets. The prototype system contains components that draw from two main disciplines namely that of text analysis and econometric and time series modelling. The novelty of the system lies in

its ability to combine methods from these two different disciplines. The system implementation is a novel contribution to the areas of content analysis and time series modelling as it combines methods and functions from each area automating a number of processing and analysis tasks. Much of the system relies on data collections from online sources and databases, and can be used with any text or financial data in an unsupervised manner. To start the system initially a user specifies a number of inputs such as data sources. A dictionary is then supplied and the system can compute the impact of news on returns for any market given these data. Due to the efforts of companies such as Quandl more information is becoming more freely available. Systems such as the one presented in this chapter benefit from the growth of these resources and from the deluge of text data being published online. Methods of collection and aggregation are an essential foundation to building useful applications. The system presented does this by leveraging freely available data sources and open source tools, platforms, and programming languages. By combining these freely available resources, the system can use content analysis and statistical modelling to determine the impact of news on financial data. Configurations that use these available resources can benefit greatly from widespread deployment of new methods providing new opportunities and avenues for various research investigations.

4

Evaluation and Case Studies

4.1 Introduction

Chapter 3 describes the proposed system developed and combines methods of content analysis and statistical estimation with the necessary functionality to automatically collect and aggregate text and financial time series. These implementations allow the impact of sentiment on asset returns to be estimated. This framework allows any unstructured text collection or corpus to be input and combined with any dictionary or glossary of terms to perform content analysis producing a time series of sentiment. This output can be aggregated with a financial time series and any number of additional time series variables to estimate a multivariate model and hypothesis test to assess any potential relationships between the variables. The following chapter presents two case studies that use and evaluate the system by estimating the impact of different news types and sources on two financial instruments traded in different markets. The case studies evaluate the functionality of the system and its ability to import different inputs and estimate different statistical models. This allows an easy implementation to examine the intricacies that differences in text and news type, source, and timing will have in each model ultimately giving a better understanding of the role that news sentiment has on returns. The system is evaluated using different text sources and financial data, to see if differences or similarities exist in the results produced by the system.

An evaluation of the system was carried out using financial data from two different markets; the equity market using the Dow Jones Industrial Average (DJIA) and the commodities market using the price of West Texas Intermediate (WTI) crude oil. The DJIA has served as a

representation of market performance and is often seen as the most widely used and recognised market index. WTI is a grade of crude oil used in benchmarking the price of crude oil and was considered the main benchmark until recent times, when the price of Brent crude oil surpassed its market price. For both instruments, news is frequently published about their performance, with the implication of events often being reported as having a direct or indirect effect on these benchmarks. A number of news sources and text types are examined for both case studies. For the equities case study these include opinion columns *Abreast of the Market* (AOTM) published in the *Wall Street Journal* and the *Lex column* from *Financial Times* (FT). Different text types and genres are also input to the system to evaluate the potential differences that various text collections can have on the explanatory power of the sentiment variable produced by the system. The second case study investigates changing the text type and source by using different sources including a corpus of *crude oil* related news sampled from all sections of the FT, articles from the *Oil & Gas* section of the FT, and message postings on the *Oildrum* blog. In this way the system is used to evaluate the differences that formal and informal news can have when constructing the sentiment proxy and its impact on returns. The oil commodities case study evaluates the content analysis method for domain text using and algorithm previously described in Chapter 3, Algorithm 1. The modelling component of the system and the different models implemented are evaluated in each case study. These include a VAR model to estimate the inter-relationships between sentiment, returns, and other market variables that might also act as a proxy for the same effects as the sentiment variable. A rolling regression model is also employed to examine the changing influence of sentiment in time. These are combined with an hypothesis test to estimate the statistical confidence in using the sentiment variable in order to explain changes in financial assets.

4.1.1 News and Financial Markets

Markets and the price of financial instruments change constantly and are said to be dynamic. This makes future predictions and the expectation of price changes difficult to surmise. Measuring price changes can be done by using time series analysis and modelling. Explaining why these price changes occur is difficult. Many interpretations and explanations for price change have been suggested, with varying degrees of success. Explanations include reactions to macroeconomic news, market speculation, and changes in investor sentiment. These interpretations, although they have been beneficial in explaining price changes, do not explain the entire process, often accounting for a small degree of change.

Researchers and analysts often use financial returns over prices in analysis for several reasons. The first is to produce a normalised series. In the long term, an asset's price can change as the underlying value changes. These changes in value can result in long term price changes and trends. The assumptions regarding the distribution of financial returns such as assuming the mean is not significantly different from zero and is uncorrelated do not adversely impact model

building [95]. Many multidimensional statistical methods require variables to be measured in a comparable metric and assume independence in the distributions of the variables. The definition of returns used here is the log difference of prices discussed in previous chapters and implemented in the transformations component of the system (Section 3.4.1).

The log returns and additional variables used are all time series. All series used consist of observations in time with weekend and holidays removed. This is the case with financial data as at these times markets are not open and data are often not quoted for many instruments. All additional time series are merged with the financial time series. Daily time series for the financial data used were obtained from different exchanges (Table 4.1) but typically data was retrieved and updated through the *Quandl* database.

Table 4.1: Available financial data used for the presented case studies and the exchanges where data were collected.

Returns	Source	Dates	Observations
DJIA Spot	NYSE	03/01/1989 - 31/12/2014	6506
DJIA Futures	CME	06/01/1998 - 24/05/2015	4357
WTI Spot	EIA	03/01/1989 - 31/12/2014	6506
WTI Futures	CME	03/01/1989 - 31/12/2014	6506

Transformations were performed on all series where necessary. This was done so as to create as close to a stationary or weakly stationary series as possible for each of the variables used. All series were tested for stationarity using the ADF test (Augmented DickeyFuller test) and every model incorporated adjusted standard errors to account for potential variance bias and heteroskedasticity. Newey-West adjusted standard errors were used for every regression model. Detailed descriptions of transformations calculated by the system have been previously described in Section 3.4.1 and further detail is give in the following sections where necessary.

4.1.2 The Impact of Sentiment on Financial Returns

For each case study the impact and explanatory power of the sentiment variable produced by the system is assessed. First a long-term VAR model is estimated with sentiment, then variations in text source and type are evaluated, a rolling regression model is then estimated to assess the impact of sentiment over time, and finally any relationships to volatility are investigated.

The equation for the VAR model for each case study is described in detail in each section. Using this model the average impact of sentiment is computed by the system. For each model the statistical significance of the coefficient computed is assessed and tabulated. One VAR model can be compared to similar model estimations using the AIC values as described in Section 2.3.2.

A hypothesis test is calculated by the system to test the statistical significance of the sentiment variable as an explanatory variable for asset returns. The hypothesis that is formulated

for each regression model estimated by the system has been defined in Equation 3.7. The hypothesis test is used to determine if there is a significant difference from the null hypothesis that sentiment does not impact returns and that the coefficients for sentiment are not significantly different from zero. The larger the chi-square value obtained from the hypothesis test then the greater the confidence in rejecting the null hypothesis. The value is compared to a chosen critical value (based on statistical confidence) and depends on the model computed. This gives more confidence towards sentiment being a statistically valid predictor of returns and rejecting the null hypothesis that it does not have an impact. This calculation is computed by the system using a hypothesis testing library in *R* (Section 3.4.3).

The adjusted-R^2 value is reported in several models to show the potential change in variance explained by each model. Low R^2 values are often typical when explaining returns [95] but important conclusions can still be made about how changes in an independent variable can be associated with changes in a dependent variable. A low value can be problematic however if a prediction is to be produced from the model as the prediction interval may not be precise enough to be useful. Precautions taken in the following model estimations to ensure results are not spurious and to address concerns of model uncertainty include: choosing variables based on previous justifications from the literature, ensuring the time series variables are consistent in the time period they represent, that outliers are not skewing results, and that the coefficients are individually or jointly significant.

Texts of different varieties: news reports and opinions, blogs, specialist news and chat messages, can contain information that is directly or indirectly related to one or more events. The texts comprise one or more types of data: (1) accurate reports of events; (2) predictions of future events or the reporting of events; (3) the exaggerated or censored versions of events; and (4) false accounts of events. Straight-forward reporting will simply reflect what happened and will be largely factual. Prediction of events or their reporting, may be factual, based for instance on a robust quantitative analysis, with some bias of analysts and a modicum of affect. Exaggerated or censored texts and false reporting has a very distant relationship with facts. This cline of texts is related to the text types given at the beginning of this section: News reportage, in principle, comprises accurate reports; editorials and the expert-written opinion columns that often comprise of predictions; blogs and comments may comprise of exaggerated or censored versions of events; and messaging services have been known to circulate false accounts which are not corrected or not verified. When looking for affect, it may be found in affect laden texts and less so in factual descriptions. Hence, previous authors who have analysed the content of news for affect have focused on opinion columns and messages [8] [42] [96]. Typically the choice of text has been a subjective choice particularity in financial literature. To test if text type and source has an influence on the sentiment variable that is generated by the system and evaluate its impact on returns, text collections consisting of different text types, such as formal or informal news, are scraped from different sources to assess potential variations in results from the content analysis method implemented.

For the crude oil case study, a section is devoted to the linguistic element of the content analysis method by incorporating domain glossaries into the method, consisting of words and phrases used in the crude oil industry. This is done by using the algorithm described in Section 3.3.2 to assess how changes in the content analysis method can influence the construction of a proxy for news sentiment. The domain glossaries are used to augment the GI dictionary and a comparison is made between the impact of the sentiment proxy created with and without the domain glossaries and its influence on the crude oil benchmark.

All markets are susceptible to wild price fluctuations, being exposed to events that are inherently extreme such as natural disasters, political turmoil, business and economic crises. The parameters (mean, variance) of asset returns are known to vary in time and may relate to these market changing events. Changes in volatility and information processing in financial markets have been widely studied and also reflect these events. Such events are recorded, reported, and predicted in news publications and disseminated to market participants by means of news outlets. It is intuitive to think the effect of sentiment may also vary in time. To investigate this effect, the system employs a rolling regression model to examine the variation in the impact of sentiment and performs a hypothesis test to assess if sentiment is still a valid explanatory variable for financial returns. Market business cycles are also included as a variable in another regression based estimation and any relationship with the sentiment variable is assessed. The final section of each case study examines potential links of asset return volatility with that of the sentiment variable and news volume. Links between the sentiment variable and asset returns volatility are made for both the case studies evaluated in this chapter using a regression based analysis.

The work presented in this chapter contributes to content analysis and the application of statistical methods in finance and the evaluation of the system described in Chapter 3. The explanatory power of sentiment on a market index is verified with an alternative source of news corroborating the importance of news type and choice of source. Opinion based articles have an impact on a financial market that is often reversed, showing a transient reaction by market participants demonstrating no new information is actually being reported. The time varying influence of sentiment is noted, like many parameters of asset returns a change is noted across time. This system is successfully applied to another benchmark in a different financial market that of commodities, where different text corpora are collected and imported into the system, a sentiment proxy is generated and aggregated with time series data for the commodities market. The use of a proxy for information in news, regardless of its interpretation, is shown to vary in predictive power in time. The use of rolling statistical methods shows the performance and impact of sentiment in different periods. Alternate periods are shown to have varying levels of predictive capability for sentiment in explaining the changes in the market and price.

4.2 Equity and Sentiment

Data used in the following case study include the Dow Jones Industrial Average (DJIA) and news related to the equity market. The DJIA is a price-weighted average of 30 blue chip stocks considered influential industry companies. A widely followed index, it is considered one of the most influential indicators of stock market performance. The performance of the DJIA is influenced by numerous factors ranging from economic and corporate performance to worldwide political events, natural disasters, and economic conditions. The DJIA and equity market has been studied widely in financial literature. Its movements have been related to news about events occurring in equity markets. This makes the DJIA and accompanying news a good choice to evaluate the system described in Chapter 2.

4.2.1 Time Series and Text Data

4.2.1.1 Financial Time Series

The summary statistics and preliminary analysis of the DJIA are presented in this section. The time series plot of the DJIA close and trading volume is illustrated in Figure 4.1 and shows the overall movement of the market from 1989 to the end of 2014. The plot shows periods of growth and decline in the daily value of the DJIA. It is clear that the DJIA movement is dynamic and suggests that a number of factors may be influencing the level of the DJIA.

The summary statistics for the spot and futures DJIA series are shown in Table 4.2 where the time period is due to the availability of data collected by the system from the Quandl database.

From Table 4.2 the unconditional mean shows a low positive average return for both series. The standard deviation of returns is similar for futures and spot returns. This measure is the unconditional standard deviation and does not account for the changing variance that is characteristic of prices and returns. Despite this, the volatility between spot and futures should be similar. Skewness and kurtosis can give an indication to the shape of the returns distribution and its symmetry. The DJIA spot series sees a smaller negative skewness as compared to futures. For futures the period where data was available shows positive skewness. Excess kurtosis is seen for all series which is typical of financial returns. Both series suggest they are leptokurtic with long tails and a high peak, an observation that has been seen in all series of daily financial returns across asset classes [95].

Correlation of variables can measure the linear dependence with the autocorrelation function (ACF) measuring linear dependence of a variable with past values of itself. The ACF is important in characterising linear effects that may be present in a time series. The ACF of log returns of financial data are typically not significantly different from zero, this is evident from the presented ACF of the series in Table 4.2. In financial literature it is assumed that the returns of an asset are not predictable and have no autocorrelation. A small degree of correlation in returns can be seen. This slight correlation has been attributed in the literature to the calculation of

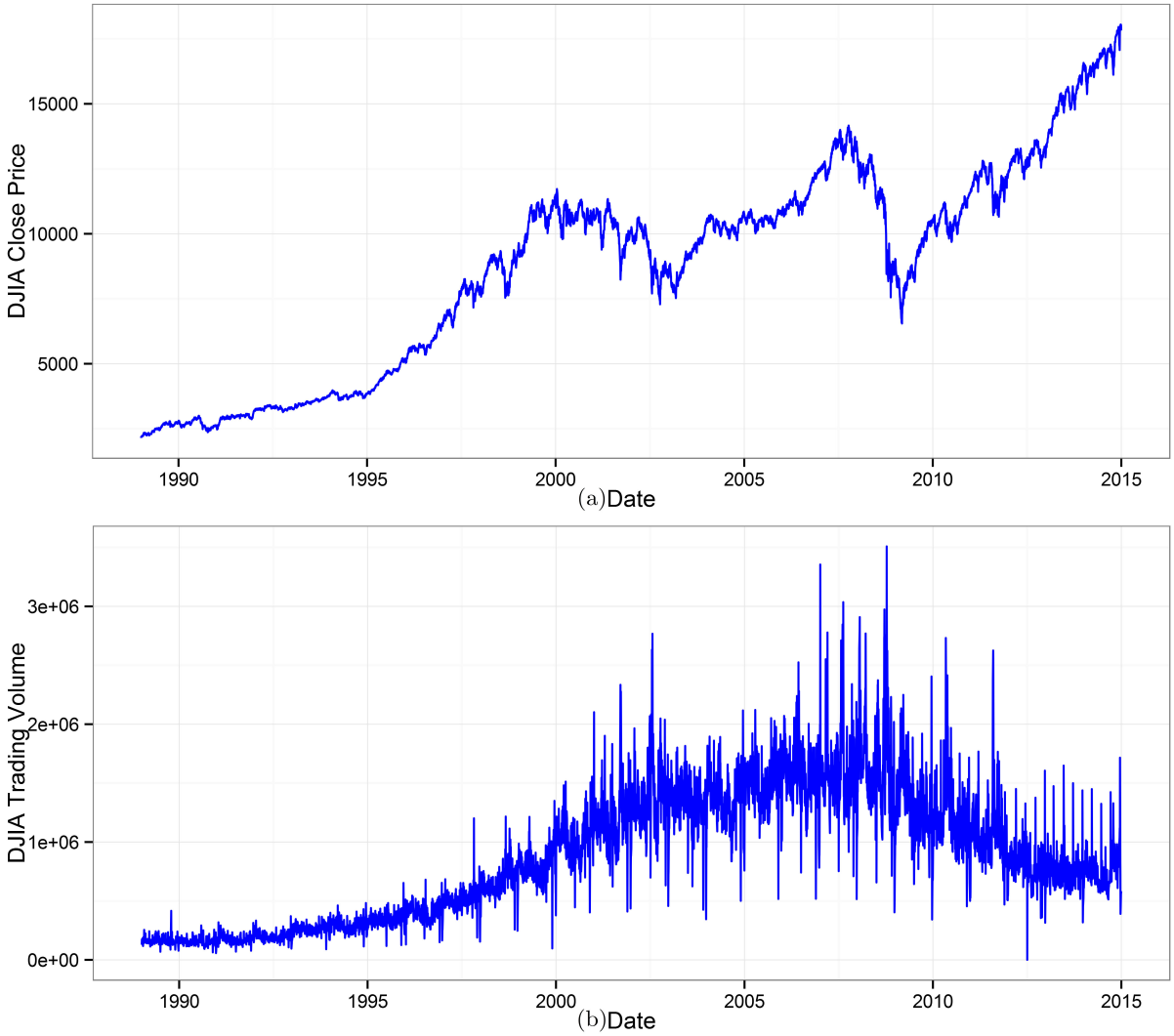

Figure 4.1: (a) Dow Jones Industrial Average daily closing index value and (b) volume traded for the period 1989-01-03 to 2014-12-31 (n = 6508)

Table 4.2: Summary statistics for time series of returns for DJIA spot and futures returns. Where μ is the mean, σ the standard deviation, δ skewness, and γ kurtosis. Values (1) to (5) are values of the autocorrelation function for five lags. The number of observations is denoted as n

Returns	$\mu(10^{-4})$	$\sigma(10^{-2})$	δ	γ	(1)	(2)	(3)	(4)	(5)	Dates	n
DJIA(S)	3.11	1.08	-0.21	8.47	-0.05	-0.04	0.02	-0.01	-0.05	03/01/1989 31/12/2014	6506
DJIA(F)	1.85	1.21	0.15	10.53	-0.07	-0.05	0.03	0.00	-0.04	06/01/1998 24/04/2015	4357

observed returns and issues surrounding how prices are determined in an exchange, such as non-synchronous trading or bid-ask bounce [95]. In an autoregressive (AR) based model the autoregressive component includes lagged variables of the dependent variable to account for these dynamics and potential multicollinearity. Although a high degree of dependence is not observed, the lagged variables account for the possibility that an observation is influenced by a previous one. The AR component allows multicollinearity to be reduced by accounting for dynamic patterns in the model by using lagged observations and no additional parameters, thereby creating a more parsimonious model. From the summary statistics presented here it is clear that the returns distribution for the DJIA has parameters that are typical of financial returns in general [95]. The DJIA index is chosen for further evaluation due to data availability and also to allow a comparison of the results produced by the system to results presented in the literature. The DJIA futures are not used in further evaluation in this chapter as less data is available overall from exchanges and online databases as compared to the spot data. This means results from the DJIA spot and futures series can not be reliably compared as the time and volume of data is different. This also means the results obtained from using the DJIA futures data cannot be easily compared with results from the literature.

4.2.1.2 Text Data

The Wall Street Journal (WSJ) is an accepted authoritative publication that reaches a wide audience and carries a column titled the *Abreast of the Market* (AOTM). The AOTM column is concerned with news regarding market activity, specifically large-capitalisation stocks quoted on the Dow Jones Industrial Average (DJIA). In the study by Tetlock [96], an attempt is made to predict daily returns with information extracted from this column. The conjecture is that information contained in the column is related to movement in the DJIA. By analysing the content of the column a time series of sentiment is created that acts as a proxy for investor sentiment that is then used to predict movements in DJIA returns.

The system developed here automatically collects a corpus of daily AOTM news columns spanning a period of twenty years (1989 - 2008). The WSJ newspaper archive on the Proquest digital database was used to create this text collection and build the corpus[1]. In the period after 2008 the *AOTM* column is published less frequently in the WSJ (according to availability in the Proquest database), dropping to a weekly frequency. For this reason and to keep the sample set and daily observations consistent, the DJIA case study using the AOTM corpus is only extended to 2008. Work presented in the following case study initially verifies the results of Tetlock using a sample corpus consisting of the AOTM column from the WSJ for the period 1989 to 1999 but without the 1987 crash period. Tetlock accounts for this crash using a dummy variable, however this study omits this period and finds it does not impact results but its inclusion means results are skewed and difficult to replicate due to the influence of this single observation, a similar

[1] http://www.proquest.com/libraries/

influence and procedure has been noted and followed in previous studies [95]. This dataset is used to verify and evaluate previously published results from the literature (Section 4.2.2).

A corpus of general news text was collected from the WSJ to act as a benchmark and comparison with the results produced by the system using the AOTM. News in this corpus consisted of general business and economic news, sampled solely from the WSJ, and tagged as being either news articles or editorials. This sample has a wider range of business focused news than the AOTM column. By using these text collections the system evaluated the distinction between a corpus of specific, opinion based news and more general business related news and their potential for creating a sentiment variable that has explanatory information for financial markets.

The Financial Times (FT) is an international daily newspaper also concerned with business and economic news whose main rival is the Wall Street Journal. The print edition has an average daily readership of 2.2 million with the FT.com website having 4.5 million active users. The print edition combined with digital circulation has seen its distribution at the highest in its history[2]. This emphasises the reach and influence the Financial Times has through its readership, particularly through a digital medium. The *Lex* column is considered to be the agenda-setting column in the FT. It features analysis and opinions about global economics and finance. The *Lex* column is the most read and valued part of the FT and is a daily feature of the FT in much the same way AOTM features in the WSJ. This reputation, high readership, and highly sought after information suggests the content of the *Lex* column may contain valuable information that could be extracted with content analysis methods. A collection of articles from the *Lex* column was scraped from the FT online archive and compiled into a corpus by the system. The corpus spans a ten year period from 2005 to 2014 inclusive, consisting of over 20,000 news items according to availability. A summary of the text collections used are summarised in Table 4.3.

Table 4.3: Full sample of text collected for each corpora used in the equity market and sentiment evaluation.

Text	Description	Article	Tokens	Period	Type
Opinion Column	*Abreast of the Market* column run in the WSJ	5466	5,663,193	03/01/1989 31/12/2008	Editorial, Op-Ed
WSJ	Wall street journal text labelled as articles and editorials from the Proquest database	36173	32,982,341	03/01/1989 31/12/2008	Editorial, Reportage
FT Lex Column	Only articles included in the Lex Column of FT	20,195	6,559,921	02/01/2005 07/11/2014	Editorial

[2]http://www.theguardian.com/media/greenslade/2013/oct/30/financialtimes-mediabusiness

4.2.1.3 Lexica

The use of a dictionary or lexicon is central to the method of content analysis implemented by the system. The GI dictionary is used by default in the system to summarise the tone of the news articles and also in the evaluation in this chapter. In the domain of finance and accounting, researchers who have used content analysis methods have focused mainly on summarising negativity and positivity in text (Table 4.4). The negative category is used here primarily to evaluate the capability of the system in extracting sentiment from text and assessing the relationship of this sentiment variable with returns.

Table 4.4: Lexical resources used in the market and commodities case studies.

Word list	Description	Tokens
GI Negative	Category contained in the General Inquirer dictionary constructed from the Harvard IV-4 dictionary, the Lasswell value dictionary.	2005
GI Positive	Category contained in the General Inquirer dictionary constructed from the Harvard IV-4 dictionary, the Lasswell value dictionary.	1637

4.2.2 The Impact of Sentiment on Equities

This section examines the relationship between the sentiment variable generated from the *Abreast of the Market* column and the Dow Jones Industrial Average using a VAR model estimated by the system. The system computes the sentiment variable using the text analysis component and incorporates this into a VAR model with DJIA returns to estimate the average impact that the sentiment variable will have on returns. The VAR model estimated in the system is defined as in Equation 4.1, which follows from similar econometric models in the literature [42] [96].

$$r_t = \phi_0 + L_5 r_{t-i} + Exog_{t-i} + L_5 BdNws_{t-i} + L_5 Vlm_{t-i} + \varepsilon_t \tag{4.1}$$

where:

r_t = is the DJIA returns
$Exog_{t-1}$ = matrix of exogenous variables, Monday and January Dummy
Vlm_{t-i} = is the log detrended trading volume
L_5 = is the back-shift operator producing five lags of the variable
ε_t = uncorrelated white-noise disturbances

Each of the financial time series was retrieved from *Quandl* or calculated by the system such as the sentiment variable and the dummy variables. A number of transformations were also performed to create a stationary or weak stationary series. Many of the transformations have been described in detail in Section 3.4.1, additional transformations are described here also. For the log detrended trading volume (Vlm), the log of the volume series is de-trended using a moving average of the log volume series for a period of 60 observations, days in the case of the daily time series. The market level trading volume is used in the form of the New York Stock Exchange (NYSE) group trading volume. To proxy for past volatility a residual series is derived from the returns series, in this case the daily returns of the Dow Jones. The residual is created by first demeaning the Dow Jones returns series, this residual value is then squared and a 60 day moving average of this squared residual is subtracted from the series to de-trend it. It is reported that a volatility index such as VIX (Chicago Board Options Exchange Market Volatility Index) can also be used instead of this volatility variable [96]. Dummy variables to account for the day-of-the week effect and January effect are also included. The control variables and these transformations described have been drawn from models defined in the literature [22] [42] [96]. These variables were chosen so as to isolate the effect of sentiment. The conjecture is that each control variable (trading volume, volatility, and dummy variables) may also capture the effects of sentiment on returns. As such, they are included in the model to see if they proxy for the same information that the sentiment variable does.

The VAR estimation computed by the system will note the variation and inter-relationships between the variables. Firstly, the returns equation, with DJIA daily returns as the dependent variable, is assumed to have an error term ε_t that is heteroskedastic across time. The assumption that the error terms are independent across the VAR equations allows each equation to be estimated using ordinary least squares (OLS) techniques. This is an interpretation that has been described in Enders [34] and followed in Tetlock [96].

The main output of the system for the VAR estimation is the coefficients of the sentiment variable defined in Equation 4.1 and the results of the hypothesis test for including the sentiment variable as a variable in the model for explaining changes in the DJIA returns. The computed coefficients of the sentiment variable describe any potential dependence that exists between the daily DJIA returns series and the sentiment proxy. The resulting coefficients for the sentiment variable can be examined both in magnitude and by their statistical significance to see the potential impact on returns. The coefficient values of the sentiment variables and statistical significance are tabulated by the modelling component of the system and can be displayed in this form as seen in Section 3.4.4. The result of the hypothesis test for the inclusion of sentiment as a useful explanatory variable ($\chi^2[NegSent]$) are also output by the modelling component. The adjusted-R^2 value is also reported ($\bar{R}^2$) for each model estimation to show the percentage of variance explained by the model. The AIC is also reported for each model so a comparison can be made between estimated models, in the case where the results for more than one model is presented, to show the change in the estimation due to the inclusion of additional variables.

The system calculates these statistics in the modelling component and have been described in detail in the model diagnostics described in Section 3.4.3.

The results presented in Table 4.5 show the coefficient values for negative sentiment on DJIA returns computed by the system for Equation 4.1. This result summarises the effect that the negative sentiment variable has on the daily returns of the Dow Jones Industrial Average. The negative sentiment series ($NegSent$) is the standardised relative count of negative affect terms occurring daily in the column *Abreast of the market*, categorised according to the GI affect dictionary. The negative sentiment series has been standardised to have zero mean and unit variance. The mean and variance of the sentiment series remain consistent over the sample series and results are not impacted by the standardisation. Standardising the data in this way allows the coefficient values of the regression to be interpreted as unit deviation changes in the distribution of negative terms occurring in the corpus of news. That is a one standard deviation (one unit) change in the frequency of negative sentiment will impact the returns by the coefficient amount estimated. These assumptions and interpretation hold and are true for all results incorporating the sentiment variable in the rest of this chapter.

Table 4.5: Coefficients for the sentiment proxy from Equation 4.1. The model was computed using data from 1989-01-03 to 1999-11-16 (n = 2717). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	AOTM	Tetlock [96]
$NegSent_{t-1}$	***-4.7***	***-4.4***
$NegSent_{t-2}$	-1.1	3.6
$NegSent_{t-3}$	1.4	-2.4
$NegSent_{t-4}$	**5.3**	**4.4**
$NegSent_{t-5}$	**4.6**	2.9
$\chi^2[NegSent]$	***16.368***	***20.8***
AIC	-18017	-
$\bar{R}^2$	3%	-

Table 4.5 shows the results computed by the system for Equation 4.1 and the results for the same equation as reported in Tetlock [96]. The results show a quantitatively similar relationship of negative sentiment with DJIA returns. A one standard deviation increase in the frequency of negative terms predicts a 4.7 basis points decrease in the Dow Jones returns on average. A reversal of this impact is also seen across the lagged variables. That is, the initial negative shock of the downward pressure on returns is temporary and is almost fully reversed. The fourth

and fifth lags are also statistically significant although positive in their influence on returns, supporting the idea of a reversal or return to mean for returns as stated in the literature [96]. The inclusion of the sentiment variable also gives a statistically significant chi-squared test ($\chi^2[NegSent] = 16.368$) supporting its contribution to the explanatory power of the model. Despite a low R^2 value the significant coefficient values can represent changes in returns from a unit change in an independent variable holding other variables constant which still provides useful information regarding the impact of sentiment. More detail regarding the interpretation of the R^2 value is in Section 4.1.2. All series are found to be stationary according to the ADF test with the variance bias and heteroskedasticity accounted for using Newey West adjusted standard errors (Appendix A, Section A.1).

The results from the literature that show the explanatory power of sentiment from news on market returns has been recreated by the system as shown in this section. The use of the system to compute these results highlights its ability to adequately collect and aggregate the necessary data and compute the models accurately. Although some discrepancy between the results presented in this thesis and the original work by Tetlock [96] exist, the differences are comparatively small. This may be partly due to issues of data aggregation or collecting the exact data in the original study, or details in the computation done by the GI program versus the content analysis component implemented here. Despite this, the similarity and verification of results show confidence in this system and serves as a successful evaluation of the method and implementation of the components of the system presented in Chapter 3.

4.2.2.1 Robustness of System Results

To test the robustness of the results presented in Table 4.5, a closer examination is made of the data samples and methods. First, the original data sample is extended. The initial sample set was selected so as to verify the influence that the sentiment proxy has on returns for the period of 1989 to 1999 keeping in line with the results in the literature and to test the functionality of the system, primarily the content analysis module and output from the model estimation. In the following section the data sample is extended to 2008 while the same VAR model is estimated using the same variables as defined by Equation 4.1. As the availability of the *Abreast of the Market* column decreases to weekly publications by the end of 2008, the study is extended to as far as there are daily observations. This ensures consistency in the dataset and a result that is more comparable to previous findings.

The results presented in Table 4.6 further corroborate the previous findings in Table 4.5 for the impact of the negative sentiment variable on returns. The results in the table show that by extending the sample set, the resulting coefficients of negative sentiment from the AOTM corpus are robust and statistically significant. Overall a slight decrease in the magnitude of the coefficients and statistical significance is observed. The hypothesis test for the inclusion of the negative sentiment as a causal variable is also still significant, although slightly reduced in

Table 4.6: Coefficients for the sentiment proxy from Equation 4.1. *AOTM* uses negative sentiment extracted from the Abreast of the Market column. The model was built using data from 1989-01-03 to 2008-08-25 (n = 4761). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	AOTM
$NegSent_{t-1}$	*-3.09*
$NegSent_{t-2}$	-2.01
$NegSent_{t-3}$	2.20
$NegSent_{t-4}$	**1.23**
$NegSent_{t-5}$	3.51
$\chi^2[NegSent]$	**11.82**
AIC	-30737
$\bar{R}^2$	2%

magnitude (11.82). A low $\bar{R}^2$ value is noted as before due to the difficulty of explaining overall variation in returns. The consistency of results demonstrates that this model can be used by the system to continually update the news and financial time series data to compute the average impact of sentiment on returns.

A stemming algorithm was applied to the text corpus to examine the potential influence that changing morphology has on the sentiment variable produced by the content analysis component. This was done so as to assess any potential differences between the content analysis component of the system developed and the original GI program. These differences relate directly to the frequency count computed by the GI program and the system implemented here. Differences in the frequency of terms may be due to additional word senses possibly being detected in the General Inquirer program as terms have been hand annotated with additional information regarding their context of use and contain hand coded rules in the GI system code for disambiguating how terms may be used in a given piece of text. The use of a stemming algorithm provides an automated and consistent method of accounting for all term occurrences regardless of morphology or tense. The Porter stemmer implementation from the NLTK library has been used by the system as described in Section 3.3.2 to perform stemming on the text here.

Stemming the text increases the overall impact and significance of the coefficients (Table 4.7). The average number of negative terms per day is increased by using the stemmed version of text corpus. It may be seen from the results that with more word senses being found, the predictive power of the negative sentiment variable may increase. Although a higher number of terms are accounted for, the chance of error may also increase as some stemmed terms may not be related

Table 4.7: Coefficients for the sentiment proxy from Equation 4.1. The Stemmed variable used a corpus of text that had been stemmed to account for more word senses. The model was built using data from 1989-01-03 to 1999-11-16 (n = 2717). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	AOTM (Non-Stemmed)	AOTM (Stemmed)	Tetlock [96]
$NegSent_{t-1}$	***-4.7***	***-6.1***	***-4.4***
$NegSent_{t-2}$	-1.1	0.6	3.6
$NegSent_{t-3}$	1.4	-0.1	-2.4
$NegSent_{t-4}$	**5.3**	***5.2***	**4.4**
$NegSent_{t-5}$	**4.6**	**4.6**	2.9
$\chi^2[NegSent]$	***16.368***	***21.856***	***20.8***
AIC	-18017	-18023	
$\bar{R}^2$	3%	3%	

to their definition in the dictionary. No noticeable difference is seen in the amount of variance explained by either model as seen in the $\bar{R}^2$ values. As it is difficult to ascertain the accuracy of the method in terms of semantic meaning and to allow the results of this evaluation to be compared to those in the literature, the unstemmed text collection is used in the evaluation of the system in this chapter.

The influence of negative sentiment derived from text on financial returns is apparent. Information extracted from the content of relevant news holds explanatory power for DJIA returns. This finding gives support for using a proxy for sentiment as an explanatory variable for returns. This proxy should be created from a collection of topic relevant text. The results presented in this section also verify the system and its ability to collect, aggregate, and analyse the text and financial data computing the impact of news on returns.

4.2.2.2 Summary

The information contained in the news is likely to be already known to the market and as such, incorporated into prices. This is intuitive as articles and columns are written at the end of a trading day before being published. Despite assumptions of whether information is known to the market or not, a response is still noted from the market from the release of news in the Wall Street Journal, detected with statistical confidence in the change in returns and can be captured using the negative sentiment variable from news. Results using the positive sentiment proxy show no significance. Theories have suggested this result may relate to markets responding

more strongly to negative news than positive. The results are not reported and no significance was shown for the positive sentiment proxy influencing DJIA returns. Positive sentiment was estimated from the corpus in the same way as negative sentiment using the GI dictionary and computing the impact from the system's modelling component. The conclusion then suggests that negative sentiment does have a significant, although temporary, impact of up to -4 basis points on the future of the Dow Jones Industrial Average returns that is subsequently corrected for. The results show a positive evaluation of the system and its ability to collect and aggregate text and financial time series data and compute a model to determine any potential relationships between the variables with a high degree of statistical confidence as seen from the individual coefficient values, their statistical significance, and the hypothesis tests carried out.

4.2.3 News Genre Variation and Equity Returns

The content of the *Abreast of the Market* column is seen to have some useful information for explaining the movements of DJIA returns. Whether this is specific to the source or type of news is examined in the following section. To perform this investigation, a more general sample of news was collected from the Wall Street Journal and a corpus constructed using the system as before. The potential influence that the genre of text can have is also assessed by examining another well received column from a rival reputable source; the Financial Times. The *Lex* column is a widely read agenda setting column for the FT. The same VAR model is estimated as in the preceding section (Equation 4.1) but two models are estimated one using the sentiment variable computed from the WSJ general business news corpus and the other from the FT *Lex* column corpus.

4.2.3.1 Text type comparison

Many of the studies in finance that have used content analysis methods to generate sentiment have used exclusively used opinion columns from major publications [96] [42]. To evaluate if a difference exists between using an opinion column and general news articles from a major publication, two different corpora were compiled to test if sentiment extracted from either has higher explanatory power for returns. The content from the AOTM column as a proxy for sentiment is compared to the content of general from news from the WSJ. Texts were collected from across the WSJ that had been tagged as news articles and editorials. In this way, content from a wider range of articles published by the WSJ is examined to see if they will give similar or different result than using the AOTM column alone to generate sentiment. Table 4.8 compares this hypothesis by comparing the negative sentiment variable computed from WSJ business news and the AOTM column.

It is seen that the coefficients for negative sentiment in the WSJ corpus are smaller in magnitude (-3.6 for the first lag) and not as statistically significant as for the AOTM column (-4.7 at 95% confidence). The hypothesis test also shows no significance for the inclusion of

Table 4.8: Coefficients for the sentiment proxy from Equation 4.1. AOTM uses negative sentiment extracted from the *Abreast of the Market* column, and *WSJ* consists of news articles and editorials sampled from all sections of the Wall Street Journal. The model was built using data from 1989-01-03 to 1999-11-16 (n = 2624). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	AOTM	WSJ
$NegSent_{t-1}$	**-4.70**	-3.60
$NegSent_{t-2}$	-1.10	-0.62
$NegSent_{t-3}$	1.40	-1.64
$NegSent_{t-4}$	**5.30**	*3.69*
$NegSent_{t-5}$	**4.60**	1.57
$\chi^2[NegSent]$	***16.37***	7.31
AIC	-18017	-17396
$\bar{R}^2$	3%	2.9%

the general news WSJ negative sentiment variable with a $\chi^2[NegSent]$ value of 7.31 versus 16.36 for the VAR model using the AOTM data. The $\bar{R}^2$ value decreases marginally by 0.1% when using the WSJ corpus. Tentative conclusions can be drawn from these results. From casual observation of the sample text from each corpus, it is seen that news in the WSJ business corpus contains more general news reporting on a number of diverse topics, while the AOTM column corpus is specific to equity markets and DJIA companies. Using the WSJ general business news corpus a sentiment variable with more noise is produced by the system. The result highlights the importance of choosing a relevant and valid corpus of text when computing a proxy for the content of news from the system.

As before, this time period sample is extended from the period of 1984 - 1999 through to 2008. The system estimates the same model as before using Equation 4.1 changing the sentiment variable in each case.

The results of the negative sentiment variable computed from the AOTM corpus show a slight decrease in the magnitude and significance of the coefficient values for sentiment but still remain statistically significant. The hypothesis test also remains significant (11.81 at 95% confidence), rejecting the null hypothesis that sentiment does not contribute to the explanatory power of the model. The hypothesis tests looks at the joint significance of the sentiment as a predictor for returns as outlined in Section 4.1.2. The WSJ however, shows no significant coefficients and the hypothesis test is not significant, indicating the null hypothesis cannot be confidently rejected and suggests that the WSJ negative sentiment proxy does not add to the explanatory power of

Table 4.9: Coefficients for the sentiment proxy from Equation 4.1. Negative sentiment was computed from the Abreast of the Market column, and *WSJ* uses the news articles and editorials corpus. The model was built using data from 1989-01-03 to 2008-08-25 (n = 4761). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	AOTM	WSJ
$NegSent_{t-1}$	*-3.09*	-1.91
$NegSent_{t-2}$	-2.01	0.83
$NegSent_{t-3}$	2.20	-2.01
$NegSent_{t-4}$	**1.23**	*2.77*
$NegSent_{t-5}$	3.51	1.86
$\chi^2[NegSent]$	**11.82**	7.50
AIC	-30737	-29740.5
$\bar{R}^2$	2%	2%

the model. Results remain insignificant for the negative sentiment proxy extracted from general WSJ news (Table 4.9). The magnitude of the coefficients for negative sentiment decreases in each case but more so for the WSJ business news corpus.

4.2.3.2 Dow Jones Industrial Average and The Lex Column

The *Lex* column corpus consists of text collected from the FT summarised in Section 4.2.1.2. The *Lex* column is a daily feature of the FT and contains opinions and analysis of global economic news. It has been shown that the choice of text from reputable sources can have an effect on the explanatory information of the sentiment variable as seen in the previous section. By using a collection of *Lex* column articles scrapped from the FT archive, the influence that an agenda setting and opinion based news column has on the DJIA returns can be further evaluated by the system.

A model is computed by the system as before using a VAR model defined by Equation 4.1, with a proxy for negative sentiment being generated from the FT *Lex* column corpus. DJIA returns are again used to evaluate the impact of the sentiment variable produced. The sample period is from 2005 to 2014 inclusive. This period was chosen due to availability of the *Lex* column news articles, where articles were downloaded from the online archive and have an availability of ten years. Articles obtained from other news databases (Proquest, Lexis Nexis) have no explicit labelling for the *Lex* column articles. To ensure consistency in the text corpus,

only articles that were available from the FT website were used. The system used a webscraper to collect all available *Lex* column articles using a query and the online search function that returned only articles from that section of the publication.

Table 4.10: Coefficients for the independent variables computed from Equation 4.1. Negative sentiment was computed from the FT *Lex* column for the period of 2005-01-03 to 2014-11-07 (n = 2409). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	DJIA				
Intercept	2.1	2.1	2.0	0.7	1.5
r_{t-1}	***-14.2***	***-14.3***	***-14.2***	-2.0	-2.1
r_{t-2}	-7.7	-7.9	-7.5	3.0	2.9
r_{t-3}	4.4	4.4	4.9	9.4	9.4
r_{t-4}	-2.3	-2.5	-2.3	6.3	6.2
r_{t-5}	-6.6	-6.7	-6.4	-8.1	-8.1
$NegSent_{t-1}$		***-8.2***	***-8.4***	***-8.5***	***-8.5***
$NegSent_{t-2}$		**5.3**	**5.3**	**5.1**	**5.1**
$NegSent_{t-3}$		0.0	0.0	0.0	0.0
$NegSent_{t-4}$		0.7	0.7	1.0	1.1
$NegSent_{t-5}$		-1.5	-1.6	-1.3	-1.3
Vlm_{t-1}			10.4	0.3	0.8
Vlm_{t-2}			8.5	6.7	6.8
Vlm_{t-3}			0.0	1.8	1.6
Vlm_{t-4}			-7.0	-4.9	-5.2
Vlm_{t-5}			12.3	18.9	18.3
vx_{t-1}				*8.5*	8.5
vx_{t-2}				-0.8	-0.8
vx_{t-3}				-3.7	-3.6
vx_{t-4}				2.5	2.5
vx_{t-5}				-6.5	-6.5
$mons$					0.2
$jans$					-7.6
$\chi^2[NegSent]$		**13.86**	**14.06**	**14.21**	**14.17**
AIC	-14532	-14536	-14529	-14548	-14545
$\bar{R}^2$	2.1%	2.4%	2.3%	3.3%	3.3%

The output from the system for estimating the model defined by Equation 4.1 has been tabulated in Table 4.10. A basic VAR model was estimated initially by the system with each additional independent variable being added incrementally and the system then re-estimates the VAR model until finally computing the complete model specified by Equation 4.1. This was done so as to note any potential changes or if confounding effects appear in the negative sentiment variable and control variables. The proxy for negative sentiment shows a statistically significant (at the 99% level) coefficient for the first lag of 8.5 basis points. The null hypothesis is also rejected with the chi-squared statistic giving a value of 14.17 for the full model estimation, indicating sentiment to be a useful explanatory variable to include in the VAR model to explain returns. The statistic is computed by the system and takes into account the number of variables of the model when computing the significance. The inclusion of sentiment is shown to increase the explanatory power of the model overall with the $\bar{R}^2$ increasing with the inclusion of sentiment. Although the value is not largely significant it does indicate that slightly more variance is being explained. The impact of the negative sentiment variable is seen to almost reverse over the lagged variables with a strong positive second lag impact. This effect is similar to the transient effect of news sentiment from the AOTM column seen in previous results and reported in Tetlock [96]. As the VAR model is built incrementally it can be seen that the statistical significance of the coefficients and chi-squared test remain statistically significant for the sentiment variable with little confounding impact from other variables. This has been suggested in the literature as a justification for isolating the impact of sentiment. The sentiment variable computed by the system acts as a good proxy for new information not captured by the other control variables in the model. A proxy for positive sentiment was also computed by the system from the *Lex* column corpus but no statistically significant impact on DJIA returns was found. The results of this section support the hypothesis that the type of news text plays an important role when constructing a proxy for sentiment.

4.2.3.3 Summary

The choice of text plays a significant role in creating a proxy for information from news. In this section an evaluation of the system was presented where different text types from different sources were used to compute different sentiment variables. These variables were input into a VAR estimation to determine the inter-relationships with DJIA returns. The *Lex* column and *Abreast of the Market* column are both well-read and respected daily features in two authoritative publications. They both discuss market level news and the implication of events to financial markets and business as a whole. In particular they contain opinions and analyses and are agenda setting columns for the WSJ and FT. A negative impact on DJIA returns of up to 4.7 basis points for the AOTM and 8.5 for the *Lex* column can be accounted for when an increase in negative news occurs. Results for the positive sentiment proxy do not show any statistical significance, which follows results reported in the literature [96]. This result is reversed across

the five lags of the sentiment variable, suggesting the impact is transitory, a response to negative news that is corrected by the market.

The choice of news is of a concern when constructing such a proxy, where readership, authority, timing of the news release, and the type of news all play a role in its impact. Larger samples of news, such as the WSJ general news corpus, which was used as a benchmark, may induce too much noise to extract a useful proxy for sentiment from text. Although it is difficult to collect a representative collection of text, an awareness of the source, genre, and topic of news can greatly increase the validity of results and quality of sentiment variable produced by the system.

4.2.4 Time Varying Effects

The moments of asset returns are known to change with time. Changes in variance have been linked with market disagreement and information processing among other theories. Changes in average returns have been linked to, and predicted by, specific days of the week, times of the year, scheduled announcements and periodic releases of information. According to traditional theories of market efficiency, anomalies such as the weekend or Monday effect once identified should disappear as their presence is common knowledge. This is also true if data mining is the result of such anomalies, although this does not appear to be the case. Changes in the distribution of returns of an asset can see changes in skewness and kurtosis if sampled in different time periods, however it is difficult to fully explain why the average returns and volatility change in time. It is important to be aware of these effects when estimating a model that incorporates returns. Although predicting price from these events and anomalies is difficult, an awareness of these effects on prices can help give a better understanding of why instruments and markets behave as they do. This awareness can better inform model building. How time variance can influence sentiment and the computation of the sentiment variable by the system is evaluated in Section 4.2.4.2.

In the following section the time varying impact of the sentiment variable is evaluated. This includes examining the changing influence and explanatory power that the sentiment variable can have on returns. It is intuitive to believe that investor or market sentiment changes in time and the response of market participants to news and information changes. The explanatory power of the sentiment variable and its changing influence on returns is the subject of study in this section and is investigated by employing a different model specification in the system developed. The modelling component of the system performs a rolling regression estimation for the input data and estimates a VAR and hypothesis test for each sub-sample period of the rolling window regression.

In the following section a method of measuring the statistical significance of the sentiment variable in time is presented using a moving window method of regression. The changing explanatory power of the sentiment variable is examined with regard to business cycles changes

using data from the National Bureau of Economic Research (NBER). Finally, a summary of the findings and comments on the evaluation of the systems rolling regression estimation is given.

4.2.4.1 Measuring Time Variance

Market timing has a large influence on how information is processed and absorbed. Previous studies in finance have shown that overpricing can occur during market peaks, and underpricing during market troughs. A glut of good or bad news can influence a pricing mechanism causing prices to deviate from fundamentals, irrational exuberance in the case of market booms, and panicked short selling in market turmoil. The predictability of sentiment may change depending on the mood or conditions of the financial market. This can be related to expansionary or recessionary economic periods. Previous studies have indicated that news content has greater impact and predicative capabilities during recessions than expansions before factoring in peoples heightened response to negative news or asymmetric response to information [42]. Sampling data at different periods in time is somewhat common place to test the consistency of results in different business cycles in finance. Work on rolling window methods for sentiment and financial data has been less common [2] [82] [97] [109]. Due to the computational approach taken here, a rolling window can be estimated relatively easily and is chosen so a more thorough investigation can be carried out on how the influence of sentiment on returns can change in time.

It is intuitive to believe that a proxy for investor sentiment or simply the content of news and market reaction to it can vary in time due to changes in business cycle, state of the economy, or public opinion. Studies in finance have noted the changing parameters of financial markets and how they alter the predictability of a model. To investigate this hypothesis a rolling regression method has been implemented in the system, which generates a series of VAR models each using 250 observations of the daily time series data. The window size (n = 250) was chosen to represent a trading year [2].

The modelling component of the system takes in the sentiment variable and financial data and computes a VAR and hypothesis test for each window of data (time period is for 250 days). The hypothesis test is run for sentiment explaining returns and alternatively for returns explaining sentiment to evaluate if returns can be related to sentiment. The resulting significance of the hypothesis test is reported for each period estimated. The window is moved one day forward each time and the VAR and hypothesis test are re-estimated. The output gives $n - 250$ number of models and results. A time series is created to show the significance of the sentiment variable as an explanatory variable (the hypothesis test) for returns and vice versa. The time series plot shows an attributed value of one for a chi-squared test that computes a result that was significant at the 90% level and a value of two for models that compute a statistically significant hypothesis test at the 95% and 99% level. The percentage of these models that return a statistically significant hypothesis test are tabulated and presented in this section.

The impact of text based sentiment has also been linked to changes in business cycles [42].

To evaluate this conjecture, the system estimates a VAR model that incorporates information about business cycle effects. A model is computed that factors in economic business cycles using data scrapped from the National Bureau of Economic Research[3] (NBER). A recession, or period of contraction, is defined by the NBER as a significant decline in economic activity and lasting a substantial period of time (number of months) and is visible in macroeconomic measures such as GDP, real income, and employment. Between an economy's peak of economic activity and trough is considered a recession while the subsequent period from the trough to the next peak is an expansionary period. These data are compiled and available freely for use, a time series has previously been compiled by the Federal Reserve Economic Data (FRED) database[4].

To estimate the effect of sentiment during a recession, a model was constructed to include a dummy variable for the NBER business cycles, following from Garcia [42], estimating the impact of the negative sentiment variable for each business cycle (Equation 4.2).

$$r_t = \phi_0 + \sum_{i=1}^{5} \phi_{1,1} r_{t-i} + (1 - D_t)(\sum_{i=1}^{5} \phi_{1,2} s_{t-i}) + (D_t)(\sum_{i=1}^{5} \phi_{2,2} s_{t-i}) + \varepsilon_t \tag{4.2}$$

where:

r_t = is the financial returns
D_t = is the dummy variable indicating recession or expansion periods
s_t = is the sentiment variable
ε_t = uncorrelated white-noise disturbances

In summary two models are implemented that incorporated information about time variance in financial data and are estimated and computed by the system: one using a rolling window to assess the differences in model estimation of sentiment and returns, the second incorporating NBER business cycles into the regression model used as before to determine potential links of sentiment with recessionary or expansionary periods in the business cycle.

4.2.4.2 The Varying Impact of Sentiment in Equity Markets

In this section the system is used to asses s the time varying influence of the sentiment variable on returns using the modelling component of the system. Three rolling regression models are computed by the system to assess differences in the statistical power of the sentiment variable computed from the three different text collections presented in this case study thus far. These include computing the negative sentiment variable from the WSJ AOTM column, WSJ business news, and the FT *Lex* column. The results in this section show three figures displaying the time series of the statistical significance of the hypothesis test for the inclusion of the negative sentiment variable in the model. Table 4.11 shows the number of significant models as a percentage of all models estimated in the rolling regression.

[3] http://www.nber.org/cycles.html
[4] http://research.stlouisfed.org/fred2/series/USRECP

Table 4.11: Rolling VAR (250 day window) model specified in Equation 4.1 for the period 1989-01-03 to 2008-08-25 (n = 4539). The significance is measured as the hypothesis test for the inclusion of the exogenous variable. The table shows the VAR estimation with columns as the independent variable and rows the dependent variable. The result shows the number of significant models as a percentage of total models estimated for the time period.

	Statistically Significant models			
	Dependent Variable			
Independent Variable	Returns	AOTM column	WSJ	Lex column
Returns	-	94%	16%	19%
AOTM column	21%	-	-	-
WSJ	15%	-	-	-
Lex column	13%	-	-	-

The results of the rolling regression show differences in the statistical confidence of the sentiment proxy in time and across variables. The negative sentiment variable is not always statistically significant in explaining changes in returns. The results tabulated in Table 4.11 show the percentage of all models that have a statistically significant hypothesis test. These models reject the null hypothesis and indicate that sentiment has statistical explanatory power for returns. The sentiment variable computed from the AOTM column corpus has a higher number of significant models than the WSJ journal at 21% and 16% of models respectively. The *Lex* column sees 19% of models as significant for the inclusion of the negative sentiment proxy. As it is a rolling VAR model, the relationship of returns on the negative sentiment proxy is also examined. Returns are a very strong predictor of the negative sentiment for the *Abreast of the Market* column, and less so for the more general WSJ business news corpus. For the *Lex* column, returns are seen to predict sentiment even less than sentiment can predict returns. In the case of the strong predictive power of returns on AOTM, this relation could be due to the fact that the column is written before the end of the day, but those writing the column may be influenced by the market mood and movements at the time of writing. The AOTM column is also better able to predict returns, using the negative proxy derived from the column, 5% more models show significance for the inclusion of sentiment as an explanatory variable (Table 4.11). A similar observation is given in Antweiler and Frank [8] where market news and articles report on market movements that have occurred with the news containing information about the markets that has already transpired, this is reflected in their media proxy.

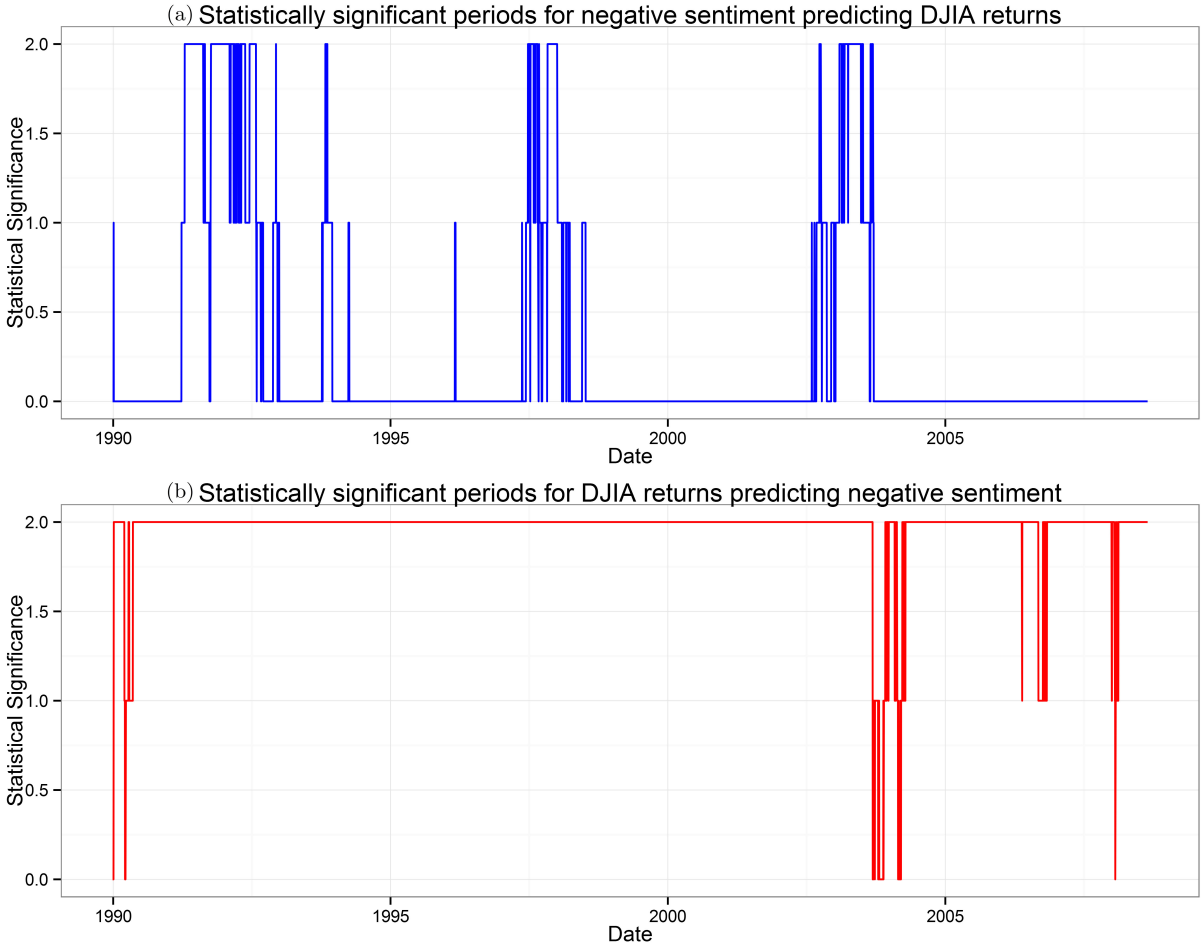

Figure 4.2: Rolling VAR (250 day window) model with DJIA returns and negative sentiment for the period 1989-01-03 to 2008-08-25 (n = 4539). Negative sentiment predicts DJIA returns (a) and returns predicting negative sentiment (b). The negative sentiment proxy was derived from the AOTM column corpus. The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

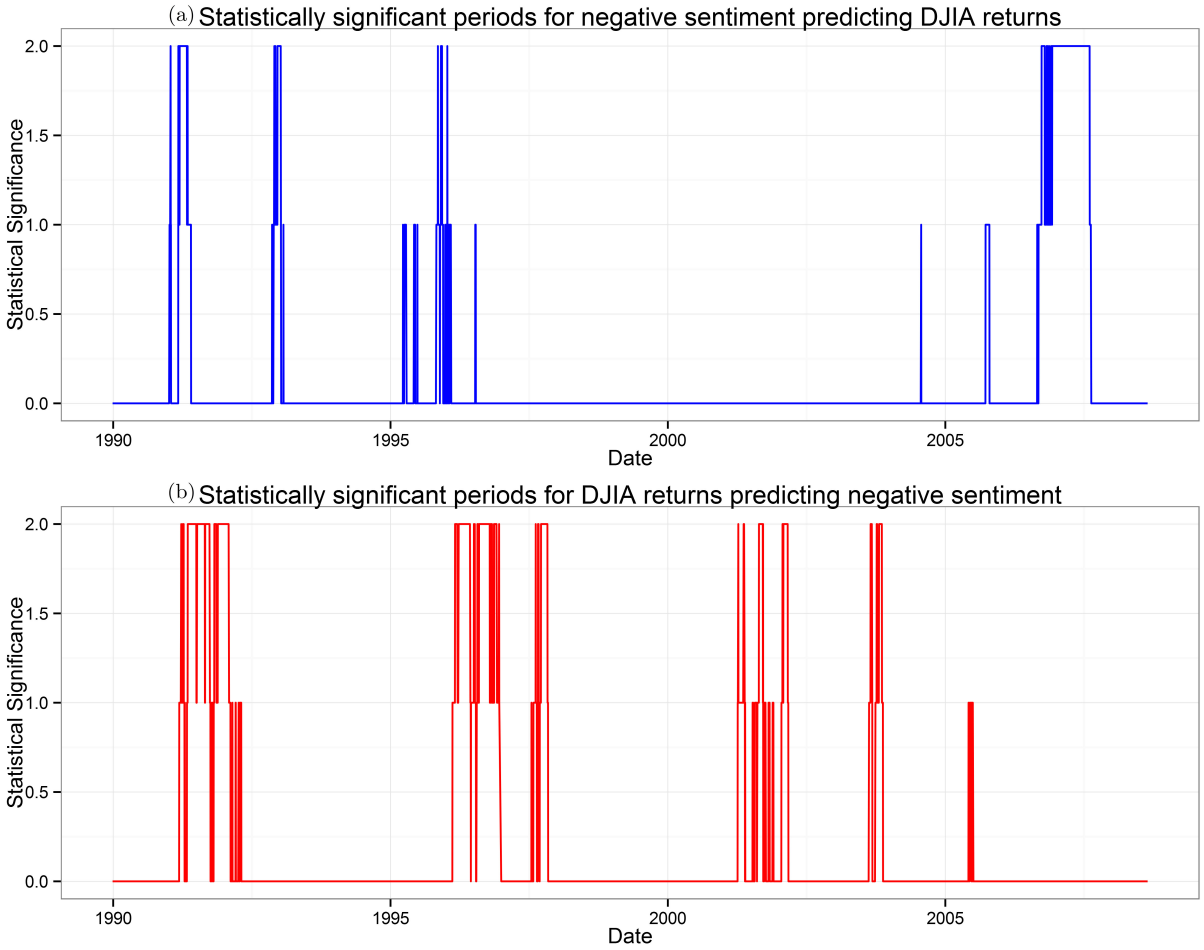

Figure 4.3: Rolling VAR (250 day window) model with DJIA returns and negative sentiment for the period 1989-01-03 to 2008-08-25 (n = 4539). Negative sentiment predicts DJIA returns (a) and returns predicting negative sentiment (b). The negative sentiment proxy was derived from the WSJ general news corpus. The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

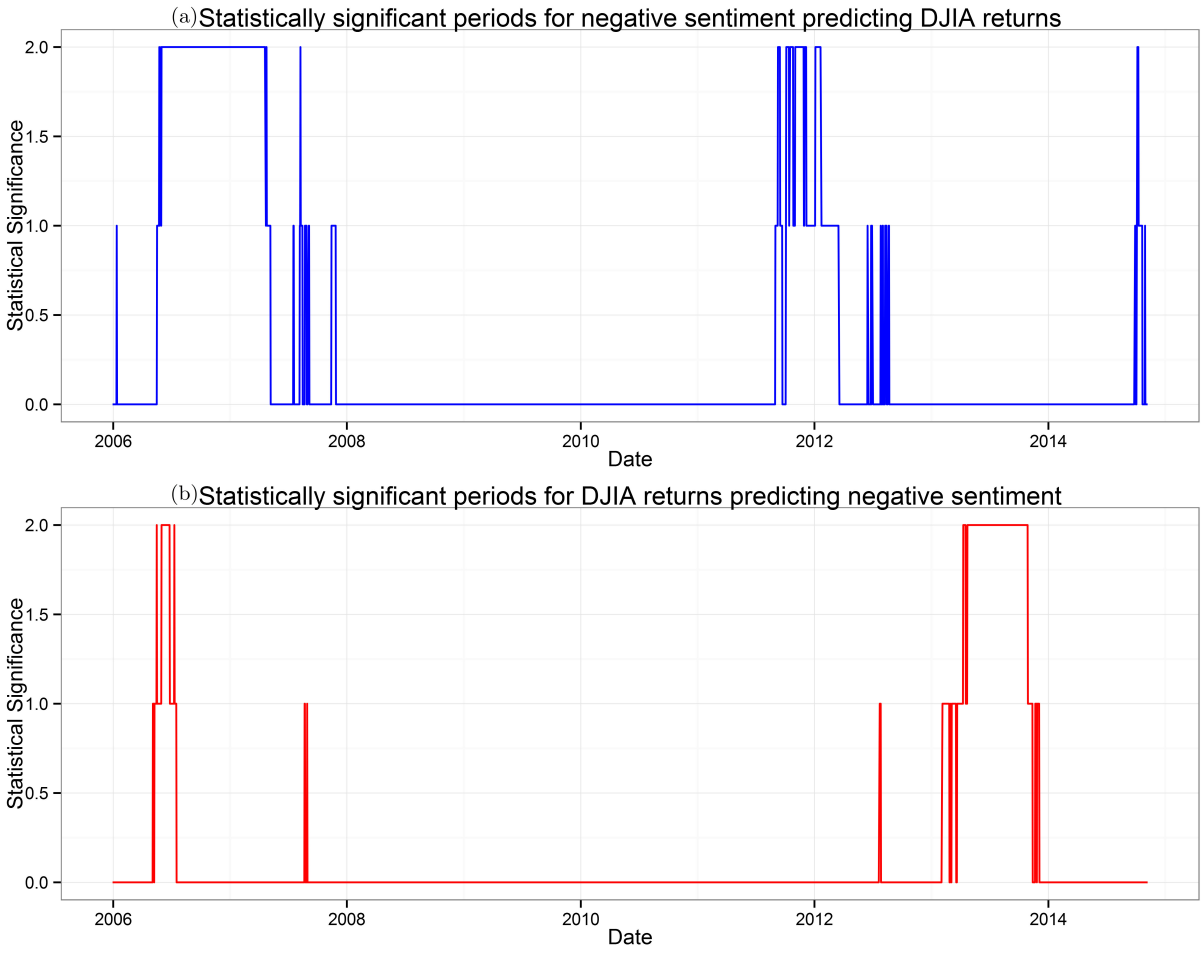

Figure 4.4: Rolling VAR (250 day window) model with DJIA returns and negative sentiment for the period 2006-01-03 to 2014-11-07 (n = 2164). Negative sentiment predicts DJIA returns (a) and returns predicting negative sentiment (b). The negative sentiment proxy was derived from the FT *Lex* column corpus. The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

Figure 4.2, Figure 4.3, and Figure 4.4 all show the changing significance of the sentiment variable in time. The results of the hypothesis test are passed through an "*if*" statement to check the level of significance returns for each model in time. The result is a time series that is seen in the preceding figures (Figure 4.2, Figure 4.3, Figure 4.4). The time series of the statistical significance of the hypothesis test computed by the system shows that the explanatory power of sentiment can vary in time. The changing influence of sentiment is apparent from the number of models that tested significant for sentiment as a useful explanatory variable. Models which do not compute a statistically significant hypothesis test for sentiment explaining returns are not invalidated due to the inclusion of sentiment but the case of sentiment as a useful explanatory

variable in the model is not as strong.

4.2.4.3 Business Cycles and Sentiment in Equity Markets

This section evaluates the results of computing sentiment with the business cycle model described in Equation 4.2. The sample period for this evaluation is from 1989 to 2008 incorporating the negative sentiment proxy from the AOTM text collection. As only one recessionary period occurs during the period that the *Lex* column sample set covers, only the AOTM text corpus is used by the system in this estimation.

Equation 4.2 contains two dummy variables for recessionary and expansionary periods. Three recessionary periods are recognised during the period of the data sample. The first recessionary period begins in January 1990 continuing to March 1991, the second period begins February 2001 to November 2001 with the last period from January 2008 lasting until the end of the data sample. The periods around the recessionary periods are considered to be expansionary periods.

Table 4.12: Coefficients for the negative sentiment variable computed from Equation 4.2. Negative sentiment was computed from the AOTM column for the period of 1989-01-03 to 2008-08-25 (n = 4789). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	Recession	Expansion
$NegSent_{t-1}$	3.59	**-3.81**
$NegSent_{t-2}$	-2.61	-2.03
$NegSent_{t-3}$	-3.36	*2.76*
$NegSent_{t-4}$	5.33	0.66
$NegSent_{t-5}$	1.88	**3.87**
$\chi^2[NegSent]$	2.719	**13.878**
AIC	-30706.14	-30706.14
$\bar{R}^2$	1.69%	1.69%

The model described by Equation 4.2, drawn from [42], incorporates both recessionary and expansionary periods by using a dummy variable to indicate the period and is computed using all data from the sample period. This allows any change in the effect of sentiment on returns in each period to be seen. The coefficients computed from the model for the negative sentiment variable during recessionary periods and expansionary periods are presented in Table 4.12. It can be seen that there is an increased effect of negative sentiment during expansionary periods than recessionary periods. The coefficient value for the sentiment variable is negative during

expansionary periods with a magnitude of 3.81 basis points. The chi-squared test statistic is also significant for this period (13.878). The coefficients for sentiment during the recessionary period are not significant. For this sample period the negative sentiment variable is seen to be more confidently associated with expansionary periods as the individual coefficient and hypothesis test statistical significance is stronger. The coefficient significance and magnitude of the sentiment variable are quantitatively similar to previous results and evaluations.

4.2.4.4 Summary

The assumptions governing returns and the movement of price in financial markets provide a foundation to construct more complex analysis and modelling. However limitations to these assumptions, such as the changing parameters that govern the distribution of returns are well-known and documented. These are important considerations when examining potential inter-relationships between variables. The preliminary results presented here build on the fact that financial markets and returns are dynamic. This leaves additional and unusual market forces open to interpretation and explanation, some of which may be explained by using hard to quantify information. This leads to the construction of a proxy for sentiment contained in news and assessing its explanatory power on the movement of asset returns. A rolling regression has been implemented in the system to examine how the contribution of negative sentiment as an explanatory variable varies in time on financial returns. The output computed by the system has been tabulated and the results of the hypothesis test represented graphically in time. The system developed allows these models to estimate the average effect that the sentiment variable has on returns and also examine how this effect changes over time. The results and output are more easily organised and visualised for interpretation by using the system (time series graph of hypothesis test significance). Models that are not significant according to the hypothesis test are not invalidated due to the inclusion of sentiment but the case of sentiment as a useful explanatory variable in the model is not as strong. In these cases returns may behave as they are assumed to and market conditions acting according to theoretical models in finance. Negative sentiment is also shown to have greater impact during expansionary periods rather than recessionary periods with the impact on returns still remaining negative. The results of this section support the idea that sentiment does have explanatory power but changes in time as do all parameters of asset returns and financial markets.

4.2.5 Equity Market Volatility and Sentiment

Some authors have made the conjecture that information contained in news may proxy for investor sentiment and this in turn can influence the behaviour of market participants [8] [96] [42]. This behaviour might induce volatility in financial markets. The volume and arrival of news impacting volatility of financial instruments has been the focus of study for a number of authors. A response to news can be detected in returns and also volatility. Some studies have noted

that a market's response to sentiment, or release of information, is different for returns and volatility [72]. In this section, the relationship between news volume, the content of news in the form of the affect proxy, and volatility of DJIA returns are examined.

4.2.5.1 Modelling Volatility

Volatility is examined using DJIA returns and the trading volume of the index. The sentiment proxy generated from the *Lex* column text corpus, and also the daily news volume for this corpus are used as independent variables. The *Lex* column is used with the GI negative category to construct the negative sentiment variable as before. The sentiment series is again standardised to have zero mean and unit variance. Trading volume is also used by the system to investigate any relationships with volatility. The trading volume for DJIA is used and transformed in the data processing component of the system by taking the log detrended (using a 60 day moving average) trading volume, the same transformation as used in Equation 4.1 throughout this section. The last variable used is the news flow, which is the number of articles published each day from the *Lex* column text corpus. The sample series is for the period from 2005 to 2014 in accordance with availability of the *Lex* column articles. The *Lex* column corpus is chosen here as opposed to the AOTM column as the news volume varies more on a daily basis with more than one article on different days, while the AOTM news volume is constant at one article being published a day.

A regression model is computed with the volatility series as the dependent variable to see if any relationship exists between volatility, the sentiment variable and news volume. Trading volume is included to act as a comparison for these news proxies and as a control variable for potential confounding effects as the literature has stated that a strong relationship is seen between trading volume and volatility. The volatility series is computed in the modelling component of the system using a GARCH(1,1) model, which outputs a daily time series of the conditional standard deviation and is used as the dependent variable in the regression. The trading volume, news volume, negative sentiment, and one previous lag of volatility are then regressed with the volatility to determine the predictive impact. This is an adaptation of a specification for a news impact function outlined in Andersen et al., [6] and implemented by Antweiler and Frank [8]. The estimation is described by the system using Equation 4.3. By using news content and volume it is possible to see if any explanatory power or relationship of these variables exists for volatility. The volatility regression Equation 4.3 is specified as follows:

$$v_t = \phi_0 + \phi_1 v_{t-1} + \phi_2 vlm_{t-1} + \phi_3 s_{t-1} + \phi_4 svlm_{t-1} + \varepsilon_t \tag{4.3}$$

where:

v_t = is a previous lag of volatility
vlm_{t-1} = is the log detrended trading volume
s_t = is the sentiment variable

$svlm_t$ = is the news volume
ε_t = uncorrelated white-noise disturbances

4.2.5.2 Volatility Regression

The resulting coefficients for estimating Equation 4.3 are presented in Table 4.13. The system computes the model and the influence of news volume and negative sentiment on volatility of DJIA returns. All independent variables have been standardised and are reported in basis points. All variables are also stationary at the 99% level, the results of the ADF test are statistically significant and reject the null hypothesis of non-stationarity (Appendix A, Section A.1, Table A.2).

Table 4.13: Coefficients estimated for Equation 4.3 for the period 2005-01-05 to 2014-11-07 (n = 2411). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	DJIA
v_{t-1}	***4228.6***
vlm_{t-1}	***206.6***
s_{t-1}	**29.3**
$svlm_{t-1}$	**35.2**
AIC	-5937.8
$\bar{R}^2$	97%

DJIA trading volume, and the lagged volatility are seen to have a large significant impact on volatility (Table 4.13). Trading volume is seen to have 206.6 basis points influence on returns. The literature and studies of volatility have stated that the trading volume of an asset is closely related to its volatility. Previous lags of volatility are also significant and have a high impact of 4228.6 basis points. Since the volatility estimation relies on the conditional values of variance a large significant impact is expected in this case, as a large $\bar{R}^2$ value is also expected as much of the variance is accounted for by this term. A statistically reliable impact of news volume and negative sentiment is seen on DJIA volatility. News volume is shown to have explanatory power for volatility, an increase in the volume of news will mean an increase in volatility, with an associated 35 basis point impact. A similar result is found in Antweiler and Frank [8] when using message board volume to predict raw volatility, however their negative sentiment metric "disagreement" does not forecast volatility. This is in contrast to the result found here where

an increase in negative sentiment leads to increased volatility of 29 basis points. Overall the results show tentative links between news volume, sentiment, and volatility.

4.2.6 Summary

The preceding section has evaluated the implementation of the system described in Chapter 3 by using financial data from the equity market and text data collected from the *Wall Street Journal* and the *Financial Times*. The results presented in this section are the output of system for determining the influence of news sentiment on equity markets by using the DJIA returns. A number of different sentiment variables are computed and passed through several models. The first VAR model looks at the average effect of negative sentiment on returns. Results for varying the text type and source as an input to the system are also presented. Following from the knowledge of changing parameters in returns and financial data, the changing impact of sentiment is also investigated. The last section of the case study sees news sentiment being linked with DJIA volatility.

The initial result and evaluation find that negative sentiment does have a significant, although temporary, impact on the returns of the market index, the DJIA. It was found that this result corroborates previous studies that suggest news sentiment has a statistically significant, albeit economically small, influence on returns. The average impact of a negative sentiment proxy on next day returns of the DJIA is apparent whether a corpus from the *Abreast of the Market* column in the WSJ or the *Lex* column from the FT is used (Table 4.14).

By extending the sample set of the study from ten years to twenty years of AOTM new articles, it is found that the impact that negative sentiment extracted from the column has on DJIA returns remains statistically significant. This result is investigated further by changing the text source and the genre of text. This was done by comparing the output of the system and changes in the models estimated when using different text corpora as an input; a general sample of business news from the WSJ and the FT *Lex* column. The development of a system to collect, aggregate, and compute the VAR model defined in Equation 4.1 allows different inputs to be quickly assessed and to remove subjectivity and potential error introduced by investigators. The system automates many of the processes of this method for computing a sentiment variable and incorporating it with financial returns in a statistical model. The evaluation of the system presented in this section show its ability to recompute results reported in the literature and are shown to be consistent with different text inputs.

The changing impact of sentiment through time is also investigated by employing the rolling window functions implemented by the system and business cycle data. It was that negative sentiment correlated more with expansionary rather than recessionary periods. Negative sentiment is seen to have less impact in terms of magnitude but remains statistically significant. Overall, returns are seen to have a strong predictive impact on sentiment. Returns will usually impact or relate to sentiment while sentiment will only sometimes impacts returns and will not be as

Table 4.14: Coefficients for the sentiment proxy from Equation 4.1. Negative sentiment was computed from the Abreast of the Market column in the Wall Street Journal (AOTM), and *Lex column* uses negative sentiment extracted from the *Lex* column in the Financial times. The model was built using data from 1989-01-03 to 1999-11-16 (n = 2717). Sentiment extracted from the FT *Lex* column was for the period of 2005-01-03 to 2014-11-07 (n = 2409).

	AOTM	*Lex* column
$NegSent_{t-1}$	**-4.70**	***-8.5***
$NegSent_{t-2}$	-1.10	**5.1**
$NegSent_{t-3}$	1.40	0.0
$NegSent_{t-4}$	**5.30**	1.1
$NegSent_{t-5}$	**4.60**	-1.3
$\chi^2[NegSent]$	***16.37***	**14.17**
$\bar{R}^2$	3%	3.3%

frequent. The day a news story is published discussion and opinions will follow along with analysis and verification of facts. The day of a particular event, online vendors and social media will have already disseminated the story, publication of these events by authoritative publications such as WSJ or FT will often come later with more discussion and opinions of the events. This is a conjecture assumed by Antweiler and Frank [8] who use message board volume to predict news stories published in the WSJ. It can be assumed from this that some of the information will have been absorbed into price already and may explain why returns are able to predict news sentiment more frequently, aside from the more well-known theory of market efficiency.

The last section of this case study has presented the results of sentiment and its relation to volatility. The regression model presented suggests sentiment and news volume can be related to market volatility. Results from this section suggest a potential avenue for future work of relating news sentiment with existing models of information processing and market volatility.

The system's framework allows a number of models to be estimated while also varying data input. Due to the models chosen for implementation, that of content analysis and VAR based models, less supervision is need to train or run the system to produce the output. The sentiment variable extracted from text is shown to have a statistically significant effect that may evolve over time depending on market conditions and volatility. The components of the system allow more models and datasets to be easily implemented and evaluated. As a result of this, the system is used on a different market and using different text data to assess whether sentiment extracted from domain text plays a role in explaining price change in the commodity market. Using the benchmark commodity West Texas Intermediate crude oil and a proxy derived from crude oil related news and sources, the system is used to aggregate and process the necessary data and estimate the impact of sentiment on commodity returns and looks at several different

model constructions to do so. The results and evaluation of which are presented in the following section.

4.3 Commodities and Sentiment

Many industries are dependent on the oil industry its influence is pervasive in many financial markets. Large movements in commodity markets have been connected to a variety of events ranging from geopolitical issues to natural disasters. The price of oil has seen extreme peaks and falls ranging from $140 a barrel to $30 in 2008 during the market downturn, and events have seen the industry benchmark change between West Texas Intermediate (WTI) and Brent crude oil. This has attracted a large number of investors to the oil commodities market and consequentially the news industry and publications have sought to cover the events that influence the price and demand for oil. This suggests that the performance of the oil benchmark (WTI) may tentatively be linked to news reporting.

The financialisation of commodity trading has meant an increase in the number of market participants in the operation of commodity markets. Due to this increased participation, the nature of information that drives commodity price formation has changed. Investors have increased trading in commodities for the purpose of portfolio diversification. Studies have shown the average returns in commodity futures are the same as equities but are negatively correlated with equities and bonds, demonstrating their usefulness for diversification [45]. Due to this financialisation in commodity markets and increased participation, increased irrational behaviour may occur from new or uninformed traders, who may make decisions on future price movements independently of supply and demand fundamentals. These distortions open the commodity market to the whims of speculation and influential news events.

Oil is and will remain for some time one of the main fuel sources of the global market. Its presence is a dominating asset in investment, influencing economic policy, development, and other financial markets. For these reasons the commodities market is a compelling one to examine the influence of news and evaluating the system's ability to collect and estimate the impact of news on changes in the market. By collecting data from the commodities market and news text about the oil industry and market, the system can reaffirm and evaluate the results computed for the equity market news and financial indices case study presented previously.

Section 4.3.1.1 presents a case study and evaluation that will follow a similar line of investigation as the previous evaluation on equity markets. Using the capabilities of the system described in Chapter 3, text and news articles describing the commodities market are collected along with commodity market financial data. Several models are then estimated and the results output. These models include a VAR estimation as before to calculate the average impact of the sentiment variable computed on WTI returns (Section 4.3.2). The influence of domain text and corpus construction is then evaluated with some attention given to domain terminology and the content analysis method implemented (Section 4.3.3). The changing influence of sentiment over

time in the oil commodity market and the rolling window methods implemented by the system are evaluated in Section 4.3.4. The final section of the chapter, Section 4.3.5, then discusses the relationship between WTI volatility and the sentiment variable and news volume.

4.3.1 Time Series and Text Data

4.3.1.1 Financial Time Series

The summary statistics and preliminary analysis of the WTI crude oil series are presented in this section. The time series plot of the WTI close and trading volume is illustrated in Figure 4.5 and shows the overall movement of the market from 1989 to the end of 2014. The plot shows periods of growth and decline in the daily value of the WTI.

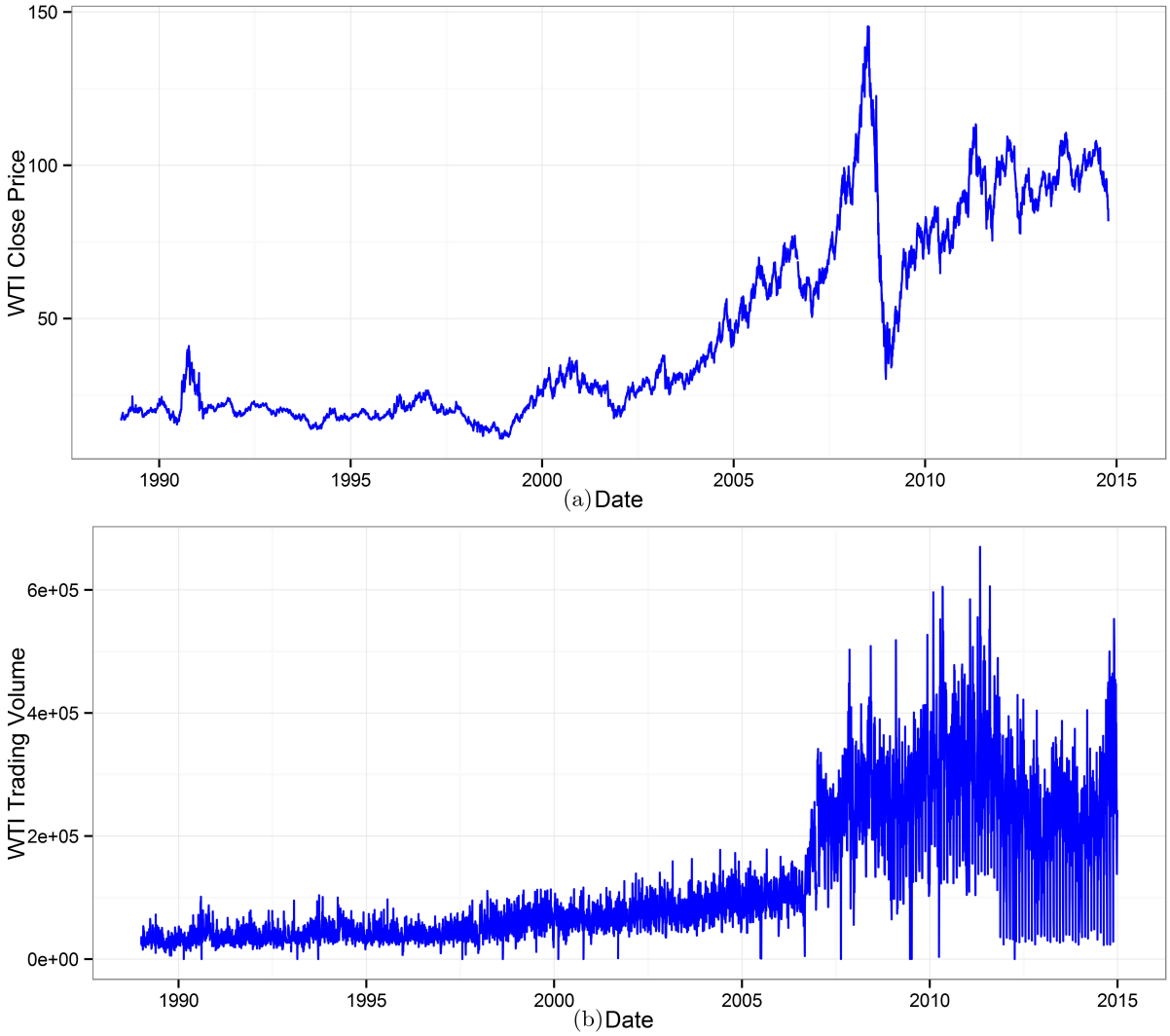

Figure 4.5: (a) WTI futures closing price in dollars and (b) volume of contracts traded (b) for the period 1989-01-03 to 2014-12-31 (n = 6508)

The 2008 market crash had a wide impact on the price of oil where the level of trading prior to this was at a peak. Commodity markets experience similar market trends to the global business market and are influenced by global economic health and business cycles but are also subject to their own market forces. These forces, such as geo-political events can directly influence the supply and demand for oil, are reported in news and reflected in the price trends seen.

Many financial instruments and assets have spot and futures prices which are highly correlated, making it difficult if not impossible to arbitrage profits. The WTI spot price represents the price of oil per barrel at any one time. The spot price also acts as the underlying commodity for pricing of the oil futures contracts. A futures contract is an agreement between participants to buy or sell an asset for a price agreed upon now at a quoted futures price, with delivery of the asset and full payment to be made at a later date. Hedgers and speculators typically trade in futures, while the latter typically does not wish to incur the cost of delivery or storage as is the case with crude oil. In some instances the futures price can exceed the spot price, making delivery of the asset now cheaper, the market is said to be in contango in this case. The converse of this is backwardation. The implication of these effects means contracts with different delivery times can be traded at different prices. Continuous contracts transition from one to another extending the delivery time, typically by one month in the case of front month contracts. Methods of smoothing adjustment can solve any numerical pricing errors from the roll-over of contracts. Front month contracts have the smallest spread between the futures price and the spot price on the underlying commodity as it has the closest expiration date. The method of price adjustment and roll over are often performed retrospectively by a data provider or exchange. The Stevens Continuous Futures premium dataset available through Quandl[5] was used as the default dataset by the system in the following evaluation to retrieve WTI futures data. The roll date of the contract is chosen as the first day of the delivery month or the expiry date if it occurs before this date. Gaps in price between contracts are smoothed using a weighted average over a period of five days where the price weighting changes over the course of the five days. The idiosyncrasies of the futures markets can be minimised using the criteria outlined above. The strong relationship (correlation) between spot and futures would suggest results from the impact of other independent variables would be similar.

Table 4.15 shows the summary statistics for the WTI spot and futures returns. The daily returns series is shown to have a small unconditional average and similar standard deviation between the two series. Both series show excess kurtosis and negative skewness. These attributes are typical of the shape of returns [95]. Little to no serial correlation is seen in either case.

A number of explanations from the market microstructure have been suggested to explain differences in spot and futures prices. Transactions and the cost of trading, bid-ask bounce, and the presence of noise traders among speculators in the futures market have been proposed. The results of the stylised facts show little obvious difference in the results of WTI spot and futures when examined for the same sample period. The unconditional mean shows a low positive

[5]https://www.quandl.com/data/SCF

Table 4.15: Summary statistics for time series of WTI spot(S) and futures(F) returns where μ is the mean, σ the standard deviation, δ skewness, and γ kurtosis. Values (1) to (5) are values of the autocorrelation function for five lags. The number of observations is denoted as n

Returns	$\mu(10^{-4})$	$\sigma(10^{-2})$	δ	γ	(1)	(2)	(3)	(4)	(5)	Dates	n
WTI(S)	2.06	2.41	-0.19	5.99	-0.01	-0.04	0.00	0.00	-0.04	03/01/1989	6506
										31/12/2014	
WTI(F)	2.08	2.33	-0.20	5.36	-0.02	-0.04	0.00	-0.01	-0.02	03/01/1989	6506
										31/12/2014	

average return. The unconditional standard deviations for both series are similar. Volatility between spot and futures should be similar. Some theories suggest the standard deviation should increase as the delivery time approaches, however no evidence of this was found in Taylor [95] when examining commodities.

4.3.1.2 Domain Text

The choice of text has been paramount in studies examining the influence of text based information and financial instruments on markets. The first case study presented in this thesis follows similar studies in the literature and focuses on the influence of news sentiment in the equity market using a market index (Section 4.2.2). The second case study presented in this section extends this line of investigation to look at the influence of sentiment in the commodities market, an area that has seen less attention in the literature. The case study in this section evaluates the system and methods for data aggregation and modelling using text and time series data from a different financial market. Table 4.16 shows the three corpora that were collected using the webscrapers in the system. Two text corpora were constructed from the Financial Times (FT) using news articles that were tagged as being relevant for the topic of crude oil and oil industry news. The last corpus was scraped from an online archive for a specialist blog, the *Oildrum* blog, that discusses the crude oil industry and catered for those trading in the oil market. The number of articles collected by the system, the number of words parsed in each text collection, the sample period used, and type of text contained in the corpus are summarised in Table 4.16.

A section of the Financial Times newspaper is dedicated to energy news and industry analysis. Three subsections exist for Mining, Utilities, and Oil and Gas. The *Oil & Gas* section consists of industry analysis and news, reporting on many facets of the oil industry such as the performance of the oil benchmarks and oil companies, exploration, supply and demand, and geopolitical analysis. A corpus of articles from this section of the FT was collected to determine the impact of industry news on the movement of the crude oil benchmark WTI. The articles were collected directly from the Financial Times online archive ensuring the correct time of publication and validity of the accompanying metadata.

Another text corpus was collected by sampling across the entire Financial Times newspaper using the keyword "crude oil". Although some repetition exists between this corpus and the *Oil & Gas* corpus, the articles were sampled from all sections of the FT newspaper. As such this corpus is determined to be more general capturing as many mentions as possible, increasing the sample while still having relevance. The articles contained in this corpus will vary from columnists providing commentary and analysis of the oil markets, news articles and overviews reporting events in the market, to the FT blog discussing commodity news. This corpus of text allows a comparison to the *Oil & Gas* corpus where the predictive power of topic specific industry news in a section of the FT and a more general news corpus of with mentions of industry news can be examined. Using sentiment from the *Oil & Gas* section of FT, it is possible to see how industry news and company performance will influence the price of WTI. General news about crude oil may only include passing reference to the industry, market, products, or companies.

The last corpus collected consists of blogs scrapped from the *Oildrum* archive[6]. The Oildrum blog ran from March 2005 to September 2013 and discussed industry analysis and the impact of energy on society as a whole. Established by a professor of political science (prof. Kyle Saunders) and a professor of mining engineering (prof. David Summers), an editorial board of respected contributors from industry and academia provided insight into direct and indirect events that would influence the oil and energy industry. The subject matter of posts ranged from distant projections to predictions and potential events that would influence the oil and energy industry from a wide range of news sources, snippets, and direct quotes. These were published in the Drumbeat posts on the Oildrum blog. This text collection would be considered informal media as opposed to formal media from major news sources like the FT. Using this corpus a comparison can be made between the different influences that formal and informal media have on the creation and explanatory power of the news proxy for commodity assets.

From Table 4.16 the volume of news is greatest for the FT all crude oil news corpus and may contain some articles that are already in the *Oil & Gas* corpus. This is apparent as the FT all news corpus can be considered as increasing the sample size of the *Oil & Gas* corpus. The corpora collected from the Financial Times have similar article lengths and word counts. The average number of articles per day is higher for the more general sample corpus. The Oildrum blogs are more varied, most being snippets and comments from those involved in the industry. As the daily relative frequency of negative terms is being used, the variation across each corpus is reduced allowing for a comparison between the corpora. The sentiment score is again computed by the system using the relative frequency of negative terms in the corpus. By doing this, changes in news volume, article length, and differences across corpora are better accounted for and are less impacted by gluts of news, sudden increases of news volume. These gluts are when a group of stories are published in a short space of time, although the volume of news will increase the frequency of negativity may not increase but relatively stay the same, this may skew results and should be taken into consideration.

[6]http://www.theoildrum.com/special/archives

Table 4.16: Full sample of text collected for each corpora used in the commodities market study.

Text	Description	Articles	Tokens	Period	Type
FT All News (key: Crude oil)	Sampled across the entire FT newspaper using the key word "crude oil"	23,157	12,252,484	04/01/2000 08/12/2014	Reportage, Editorial
FT Specialist news	Only articles included in the Oil & Gas section of FT	7,859	4,397,120	31/01/2008 03/11/2014	Reportage
Oildrum Blogs	Snippets, quotations, and excerpts featured as posts and commentary on the Oildrum blog archives	80766 excerpts	5,838,633	26/01/2007 01/09/2013	Blogs

4.3.1.3 Lexica

Authors have cautioned that a misclassification of terms may occur due to word lists, such as that used in the General Inquirer, not being fit for analysing domain specific text such as finance and business news. Filtering the GI negative and positive word lists using domain terminology for the oil industry may help reduce misclassification for domain level polarity. A method and implementation of accounting for domain terms was described in Section 3.3.2. This implementation allows the content analysis method to account for $n-gram$ level domain phrases and words aiding classification by adding information about the context of terms used in the domain text. The method used to calculate the relative frequency of negative sentiment in domain text was summarised in Algorithm 1 in Chapter 3.

Filtering the GI negative and positive terms lists using domain terminology, words, or phrases for the oil industry may help reduce misinterpretations when determining whether a term is negative or positive in the context of oil specific news. The content analysis component of the system presented in Chapter 4 uses an implementation of Algorithm 1 and is evaluated in this case study (Section 4.3.3) by incorporating two glossaries of domain words and phrases with the GI dictionary. The first of the two oil glossaries used is from Platts[7] and includes common terms and abbreviations from the oil and energy industries. The glossary contains 704 terms with expanded abbreviations. This glossary is used to account for domain words that occur in oil related news that may be misinterpreted as negative terms by the GI dictionary. As an example, it is seen that 15 terms from the GI negative category occur in the Platts glossary, and

[7] wwww.platts.com/glossary

24 GI positive terms from the GI dictionary also occur as words and in phrases in the Platts glossary. These terms include "sour gas", "open outcry", "dirty power", "availability factor", "prime mover", and "power purchase agreement". The count of negative and positive words will not be altered if GI negative and positive terms occur independently of phrases from the domain glossaries. If the term "crude" occurs separately from the compound term "crude oil", the latter will not increase the negative word count but the former will still be accounted for in the negative term count. A glossary of words from the Oil and Gas UK is also added to the Platts glossary to add more domain phrases, a summary of the word lists, tokens and detail description of these lexica are shown in Table 4.17. These are combined with the GI dictionary which acts as the base dictionary computing the frequency of negative terms.

Table 4.17: Lexical resources used in the market and commodities case studies.

Word list	Description	Tokens
Platts	Glossary from Platts who provide information and price and estimation on industry benchmarks in the physical energy markets	704
Oil and Gas UK	Glossary from the Oil & Gas UK a leading representative body for the UK offshore oil and gas industry	126

4.3.2 The Impact of Sentiment on Commodities

The relationship between the sentiment variable generated from crude oil industry news collected by the system from the Financial Times on the returns of WTI futures is examined in this section. The system computes the sentiment variable as before using the content analysis method and uses a VAR model to estimate the average impact the sentiment variable will have on the returns (Equation 4.4). WTI futures are chosen as the dependent variable as market participants and traders will typically trade futures contracts not the physical commodity which spot prices represent and requires physical storage after purchase. Market conditions specific to the futures market, such as contango and backwardation outlined before in Section 4.3.1.1, are adjusted for by data providers retrospectively by using price adjustment mechanisms. The WTI futures price is used in the following regression study. The model defined by Equation

The VAR model estimated by the system to show the inter-relationships between sentiment and returns is described in Equation 4.4 with additional independent variables included, acting as control variables, to account for potential variables that may proxy for the same information as the sentiment variable. This is a similar model and interpretation as defined in the evaluation presented in the previous section (Equation 4.1). As before, the system estimates a baseline AR

model consisting of just financial returns and incrementally adds the independent variables to assess if the addition of any of the variables interacts or has a confounding effect with the sentiment variable. The VAR model used in this section is defined as:

$$r_t = \phi_0 + \sum_{i=1}^{5} \phi_{i,1} r_{t-i} + \sum_{i=1}^{5} \phi_{i,2} s_{t-i} + \sum_{i=1}^{5} \phi_{i,3} v_{t-i} + \sum_{i=1}^{5} \phi_{i,4} o_{t-i} + \sum_{i=1}^{5} \phi_{i,5} Exog_t + \varepsilon_t \quad (4.4)$$

where:

r_t = is the WTI returns
s_t = is the affect proxy
v_t = is the log detrended trading volume
o_t = is VIX returns
$Exog_t$ = is the matrix of exogenous variables
ε_t = uncorrelated white-noise disturbances

The affect proxy s_t is again the negative sentiment extracted from the text collection that is imported to the system. To highlight this in the results that follow $NegSent$ is used as the label when reporting the coefficients values computed for the sentiment variable. Although this estimation resembles the model in the previous evaluation (Equation 4.1), the variables used differ while presentation and interpretation of the estimated coefficients and model remain the same (basis points for model coefficients and model diagnostics are the same). Each of the financial time series were retrieved from *Quandl* or calculated by the system such as the sentiment variable and the dummy variables. The transformations for each variable have been described in more detail in Section 3.4.1 and further elaborated in Section 4.2.2. The trading volume for WTI is included (v_t). The VIX index is used for the volatility measure (o_t). This index acts as a market level volatility measure, which is a widely used and popular measure of volatility. VIX measures expected volatility of the S&P500 index over the coming thirty days. The OVX (Oil VIX) would be a better proxy, being specific to commodity markets, however due to data availability does not suit some of the data samples chosen here, this series begins from 2007 onwards. A dummy variable is included for the day of the week (Monday), and month of January to account for potential weekend and business cycle effects. These control variables attempt to account for seasonal effects. Five lags of the control variables are included where specified to account for roughly a week of trading. Previous values of returns (five lags) are also included this is to account for larger than expected past return values.

News was collected from 2000 to 2014 as summarised in Table 4.16, where articles were sampled from the Financial Times that were tagged as being relevant for the term *crude oil*. This corpus was imported into the system and the negative sentiment variable was calculated from the text by using a combination of the GI dictionary and refining it using the oil glossary from Platts and Oil and Gas UK as described by Algorithm 1 in Chapter 3. The negative category

from the GI acts as a base dictionary. The word lists in the glossaries are used to ensure the negative category from GI is not counting words or phrases that are considered domain specific phrases or words and therefore may have no specific negativity or polarity associated with it. The relative frequency of negative sentiment is computed and passed as the sentiment variable to the modelling component of the system. The results of the VAR model and impact of sentiment are shown in Table 4.18.

Table 4.18 shows the results computed by the system for estimating Equation 4.4. The model is estimated incrementally as before with each independent variable being added and the model re-estimated to observe possible confounding effects with sentiment. The result shows that negative sentiment has a negative relationship with WTI futures returns accounting for an 8 basis point impact on returns on average for the period of the sample set. Much of the impact of sentiment is dispersed during the week, across the lag values. As previous studies have shown, the timing of news plays a role in the predictability and usefulness of sentiment derived from news content. Agenda setting columns such as the *Abreast of the Market* and the FT's *Lex column* typically have timely and succinct information and are often self-contained. Articles in the oil related text corpus used in the evaluation in this section contain news and commentary about events that are on-going and follow these events accordingly. The inclusion of the sentiment variable gives a statistically significant chi-squared test ($\chi^2[NegSent] = 14.322$) supporting its contribution to the explanatory power of the model. The adjusted r-squared value ($\bar{R}^2$) is also seen to increase marginally with the inclusion of the sentiment variable explaining more variance in returns (0.6% increase).

4.3.2.1 Robustness of System Results

To test the robustness of the results computed by the system presented in Table 4.18 and if any difference in using WTI futures or spot returns as the dependent variable exists, an evaluation is made using the WTI spot series as input to the system with the sentiment variable computed as in the previous section. Similarities in the summary statistic and correlations would suggest the impact of sentiment on WTI futures and spot returns would be similar. To test this and the robustness of the coefficients for the negative sentiment, the model defined by Equation 4.4 is estimated again with WTI spot returns as the dependent variable.

Table 4.18: Coefficients for the independent variables computed from Equation 4.4. Negative sentiment was computed from the FT crude oil corpus for the period of 2000-01-04 to 2014-12-08 (n = 3553). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	WTI				
Intercept	1.6	1.7	1.6	1.6	3.3
r_{t-1}	-5.2	-6.0	-6.2	-8.4	-8.4
r_{t-2}	-1.3	-2.0	-2.1	-1.9	-1.9
r_{t-3}	4.2	3.2	3.0	4.7	4.4
r_{t-4}	-3.0	-4.2	-4.4	-3.7	-3.9
r_{t-5}	-7.8	-8.8	*-9.1*	-8.7	-8.9
$NegSent_{t-1}$		-2.1	-2.3	-2.9	-2.9
$NegSent_{t-2}$		**-8.5**	**-8.6**	**-8.5**	**-8.5**
$NegSent_{t-3}$		*-8.5*	*-8.6*	*-8.0*	*-7.9*
$NegSent_{t-4}$		*-6.0*	*-6.1*	-5.7	-5.8
$NegSent_{t-5}$		1.0	1.0	0.9	0.7
v_{t-1}			-1.5	-1.3	-1.7
v_{t-2}			-0.8	-1.8	-1.4
v_{t-3}			-5.3	-5.6	-5.1
v_{t-4}			-3.2	-3.2	-3.6
v_{t-5}			1.8	1.5	1.1
o_{t-1}				**-12.9**	***-13.2***
o_{t-2}				-6.5	-6.4
o_{t-3}				7.3	7.2
o_{t-4}				4.6	4.4
o_{t-5}				3.2	3.7
mons					-16.1
jans					17.0
$\chi^2[NegSent]$		***14.9***	***15.231***	***14.479***	***14.322***
AIC	-16525	-16530	-16524	-16532	-16532
$\bar{R}^2$	0.1%	0.3%	0.3%	0.7%	0.7%

Table 4.19: Coefficients for the independent variables computed from Equation 4.4 with WTI spot returns as the dependent variable. Negative sentiment was computed from the FT crude oil corpus for the period of 2000-01-04 to 2014-12-08 (n = 3553). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	WTI				
Intercept	1.0	1.0	1.3	1.3	3.0
r_{t-1}	-9.0	-9.9	-10.0	*-12.1*	*-12.1*
r_{t-2}	0.9	0.0	-0.1	0.1	0.1
r_{t-3}	*12.0*	*10.9*	***10.8***	**12.7**	**12.4**
r_{t-4}	-4.4	-5.5	-5.6	-5.2	-5.4
r_{t-5}	-9.3	-10.1	-10.3	-10.2	-10.3
$NegSent_{t-1}$		-5.1	-5.3	-5.9	-5.9
$NegSent_{t-2}$		**-9.5**	**-9.6**	**-9.5**	**-9.4**
$NegSent_{t-3}$		**-9.0**	**-9.0**	*-8.4*	*-8.2*
$NegSent_{t-4}$		-1.9	-2.0	-1.7	-1.8
$NegSent_{t-5}$		0.7	0.7	0.6	0.3
v_{t-1}			-10.0	-9.6	-10.7
v_{t-2}			8.0	5.4	6.4
v_{t-3}			-9.0	-9.2	-8.0
v_{t-4}			-14.2	-13.5	-14.6
v_{t-5}			1.5	0.2	-0.7
o_{t-1}				***-12.3***	***-12.6***
o_{t-2}				-6.6	-6.5
o_{t-3}				*8.9*	*8.7*
o_{t-4}				4.2	4.0
o_{t-5}				0.3	0.8
mons					*-16.6*
jans					17.5
$\chi^2[NegSent]$		***15.261***	***15.551***	***15.124***	***14.9***
AIC	-16338	-16343	-16337	-16344	-16344.0
$Adjusted - \bar{R}^2$	0.4%	0.7%	0.7%	1.0%	1.1%

Table 4.19 shows the results computed by the system for estimating Equation 4.4 with WTI spot returns. On average a one standard deviation change in negative sentiment will result in a 9.5 and 9 basis point impact on returns for the second and third lagged period. The impact of sentiment is again dispersed across the five lags. The inclusion of the sentiment variable gives a statistically significant result for the hypothesis test ($\chi^2[NegSent] = 14.9$) supporting its contribution to the explanatory power of the model. The coefficients computed by the system highlight the link that the negative sentiment variable has on the movement of the WTI series. Overall it suggests there is information contained in relevant news text that can be summarised using content analysis and the system framework presented in Chapter 3. The sentiment variable produced has explanatory information that can help explain price changes in the oil commodities market, with no obvious difference existing between spot or futures commodities market. All series used are found to be stationary according to the ADF test (Appendix B, Section B.1, Table B.1)

Large outliers can have adverse effects on statistical approximations that do not use robust methods. Large tails are typical in financial data and can often contain meaningful information, with event studies sometimes focusing on large infrequent changes that may otherwise be considered outliers. Due to this, robust statistical methods are often required to check the consistency of results. Returns and negative sentiment can be winsorised to assess whether a few individual outliers are influencing results or if results are robust to their presence. A simple trimming method can remove any outliers according to criteria such as removing values in the 1% of either tail of a distribution. A more reliable method that keeps the number of observations consistent would be to weight observations in a similar method of how outliers would be re-weighted in some advanced regression models. The median absolute deviation is calculated for the sample data set. Any observations that are found to be a number of standard deviations from this central median point (3 or 5 standard deviations) can be pulled closer to the center of the distribution.

Two VAR models are computed by the system and compared to show the impact of winsorising the returns and negative sentiment data. In the transformations component of the system the returns and sentiment series are winsorised so that observations that are in the 1% most extreme ends of distributions are omitted, the tails of the distributions for returns and sentiment. The results are shown in Table 4.20. Winsorising the data does not adversely impact the significance or magnitude of the negative sentiment proxy. A slight decrease is seen in the magnitude of the leading coefficient of negative sentiment, while the second significant coefficient is no longer considered significant. The inclusion of the sentiment variable and chi-squared test ($\chi^2[NegSent] = 10.94$) is still significant supporting its contribution to the explanatory power of the model.

Table 4.20: Coefficients for negative sentiment on winsorised and non-winsorised WTI returns. The model estimated is a VAR(5) with WTI returns and negative sentiment. Negative sentiment was computed from the FT crude oil corpus for the period of 2000-01-04 to 2014-12-08 (n = 3553). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	Returns	Winsorised Returns
Intercept	1.65	2.95
r_1	-5.99	-5.05
r_2	-2.01	0.17
r_3	3.19	3.74
r_4	-4.18	-0.86
r_5	-8.83	-5.14
$NegSent_1$	-2.13	-1.27
$NegSent_2$	**-8.53**	*-7.31*
$NegSent_3$	*-8.49*	-5.43
$NegSent_4$	*-6.02*	*-5.88*
$NegSent_5$	0.95	-0.08
$\chi^2[NegSent]$	**14.9**	*10.94*
AIC	-16530	-17295
$\bar{R}^2$	0.3%	0.2%

4.3.2.2 Summary

The results and models computed by the system indicate that the negative sentiment variable has explanatory power for the returns of a financial asset. The control variables in the model estimated show little to no confounding relationship with the sentiment proxy. This indicates that the sentiment proxy is a potentially unique explanatory variable for financial returns in the commodity markets. One potential consideration of the results presented in this section may concern the economic theory underlying the model used to explain the changes in the WTI prices. The model computed by the system relies less on economic theories surrounding commodity price movements and variables that influence commodity markets but instead attempts to determine if the sentiment variable computed by the system has statistical power for explaining changes in the price of a commodity asset. As previous studies in economics and information systems have shown, many inputs, model permutations, and exogenous effects have been used to define the fluctuations of oil prices with little significance or explanatory contribution to the model [39]

[58] [104]. Despite this, the contribution of a proxy variable, particularly in a VAR framework, has explanatory power. This is particularly true for individual coefficient effects. Sentiment is shown to be robust in its effect and statistically significant. Hypothesis tests conducted also verify that the negative sentiment variable computed by the system from oil related news is a statistically valid variable for explaining changes in returns in the commodity market. For the data sample collected by the system, it is seen that negative sentiment expressed in news has a delayed effect on daily returns. This result addresses the hypothesis of whether sentiment derived from text influences other asset classes and illustrates a good evaluation of the system.

4.3.3 News Genre Variation and Domain Text

Having shown the importance that the type of news has on the creation of the sentiment variable by the system in the previous case study (Section 4.2.3), this investigation is extended further to look at the influence that different text types and genres can have on results computed by the system.

The previous section has shown an evaluation of the system described in Chapter 3 and its ability to compute a sentiment variable for a different financial market using different data and domain news. In this section variation in the news source is examined by using different text corpora as inputs and computing the average impact of the sentiment variable produced using the VAR model as in the preceding section. To perform this evaluation different news types are compared. This is done by comparing a corpus of formal news (Financial Times) to a corpus of informal news (Oildrum blogs). The three corpora used were described in Table 4.16. Articles for these text collections were chosen so the content would be domain specific, discussing topics relevant for the oil industry and commodities market. Within an authoritative news source, a comparison is made between the sentiment variable produced from a general sample of relevant news and a proxy produced from a smaller more restricted sample containing only a specific section of the newspaper. The *Oildrum* blog corpus allows a comparison to be made between the negative sentiment variable produced by the system from formal news versus informal news.

Another possible difference may exist in the dictionary used to summarise the tone of the news articles. This difference is in the interpretation of what would be considered negative and positive terms in a domain text. A word list or lexicon that contains domain knowledge may more accurately represent or extract information from text that is domain or topic specific. The evaluation in this section also examines differences in the sentiment variable computed by the system using just the GI dictionary and the combined GI dictionary and domain glossary. As before, the negative category from the GI dictionary is used and filtered using the oil glossaries; the Platts glossary and the Oil and Gas UK glossary. This evaluation looks at how the sentiment variable produced by the Algorithm 1 implemented in the content analysis component of the system can differ depending on whether domain terminology is taken into account.

In the following section, an investigation is made into how the choice of text and type of news

can influence the construction of the sentiment proxy for commodity markets. Two collections of text are built from the Financial Times according to different sampling methods. Both corpora are domain specific, both contain news articles that are topical and related to the oil industry. The last text collection was obtained from the *Oildrum* blog archive. A different text source and type, this collection allows an examination into the influence blogs and informal news may have on creating a proxy for sentiment. Lastly how sentiment is extracted from domain text is considered in more depth. As some terms may be misinterpreted with respect to specialist areas of discussion, how the sentiment proxy is constructed when domain terminology is considered is also investigated.

4.3.3.1 Evaluation of Text Type and Lexica

To evaluate the text type, the three different oil related text collections are imported into the system and used first in conjunction with the GI negative word list and then with the filtered version of this word list using the domain terminology as outlined in Algorithm 1 in Chapter 3. This evaluation compares how differences in text type and lexica can influence the explanatory power of the sentiment variable produced by the system for the WTI benchmark. The results in Table 4.21 show the coefficient values computed from Equation 4.4 for the negative sentiment variable computed using the GI dictionary and the GI dictionary filtered with the oil glossaries.

Table 4.21: The evaluation of Algorithm 1 described in Section 3.3.2 using the three corpora described in Table 4.16 and lexica in Table 4.17. Coefficients for the negative sentiment variables are computed from Equation 4.4. The number of observations used were 1587, 1416, 1182, for periods 2007-01-10 to 2014-10-31, 2008-02-29 to 2014-07-08, and 2007-01-26 to 2013-08-30 respectively. Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	GI Negative			GI negative Filtered		
	FT Crude	Oil Gas	Oil Drum	FT Crude	Oil Gas	Oil Drum
$NegSent_{t-1}$	1.92	-1.00	13.46	2.80	-2.89	4.48
$NegSent_{t-2}$	-0.04	-2.65	-2.35	-0.25	-0.11	-8.08
$NegSent_{t-3}$	-1.08	4.06	-14.33	-2.22	3.11	*-20.26*
$NegSent_{t-4}$	-2.89	-3.73	-2.03	-5.25	-1.19	-1.74
$NegSent_{t-5}$	***-8.82***	**-8.72**	*-16.68*	**-7.85**	*-7.83*	**-21.51**
$\chi^2[NegSent])$	*10.1*	*9.4*	5.9	8.8	4.6	*9.3*
AIC	4605.2	4074.4	2484.7	4606.5	4079.3	2481.2
$\bar{R}^2$	1.3%	1.4%	1.7%	1.2%	0.9%	2.1%

Table 4.21 shows the results computed by the system for estimating Equation 4.4 with WTI future returns and the negative sentiment variable computed using the GI dictionary and the combined GI dictionary and oil glossaries. Using the oil glossaries in conjunction with the GI dictionary sees a marginal decrease in significance and magnitude of the coefficients for the negative sentiment variable computed from the FT *crude oil* and Oil & Gas corpus. The use of the oil glossary has a more noticeable difference when computing the sentiment variable from the Oildrum blogs, where an increase in the significance and magnitude of the sentiment coefficients is seen. The hypothesis test for the inclusion of sentiment also changes to be significant.

By using the oil glossary and computing the relative frequency of negative sentiment using Algorithm 1, a decrease in the frequency of negative terms is seen. For the FT crude oil corpus the frequency of negative terms decreases by 24% (from 113556 to 85795), for the Oil & Gas corpus an 18% decrease occurs (106095 to 87174) and the *Oil drum* blogs sees the smallest decrease in negative sentiment at 16% (169904 to 143357).

News published in authoritative newspapers is written to be understood by a wide audience as the readership is large and varied. It is seen that the GI negative category combined with the domain glossary slightly decreases the impact that the negative sentiment series has on predicting WTI returns. The statistical significance and overall explanatory power of the sentiment variable computed using the negative category with domain word lists decreases. For the Oildrum blogs, the coefficient for the negative sentiment variable increases as does its statistical significance after filtering for domain terms and phrases. This may be due to the higher number of domain specific phrases and words being present in the Oildrum postings. Typically the blog posts on the Oildrum tend to be more specialist discussing aspects of the industry in greater depth, using more domain specific words and language.

When examining the influence of news on asset returns, it is clear that the choice of text genre and topic plays an important role in the corpus construction and the usefulness of the sentiment variable computed from this text. The use of a glossary containing domain phrases and words is useful when examining informal news, but makes less of a contribution for formal news and major publications where reaching a wider audience and readership is more important, the text as a result contains less domain specific terminology.

4.3.3.2 Summary

This section has examined how varying the news type and lexica can influence the sentiment variable computed by the system. It was found that the negative category from the GI dictionary works adequately and is not adversely impacted by misinterpretations of domain phrases and words deemed to be negative according to the GI dictionary using the content analysis method implemented. A difference is seen however when using a corpus of informal domain specific news. This may be due to articles that occur in the Financial Times being written to appeal to a wide audience whereas posts that appear on the Oildrum blog discuss details of the oil industry

and commodity market typically using more domain specific language. Improvements can be made to a basic dictionary based method of content analysis by using the algorithm described in Section 3.3.2 Algorithm 1 when analysis text that contains domain terminology. The basic negative category from the GI dictionary however may be the most reliable choice as its simple implementation means fewer assumptions and lexical resources are required for the method and results are more easily interpreted and replicated.

4.3.4 Time Varying Effects

The time varying impact of the sentiment variable computed by the system is investigated for the commodity market case study in the same way as before using a rolling regression estimation and hypothesis test defined in Section 4.2.4.1. The rolling regression model incorporates WTI returns and negative sentiment defined in Equation 4.4. The corpus of text used as the input to the system in this evaluation is the FT crude oil corpus as it is the longest time series sample of the oil related corpora collected by the system and allows more observations to be used than the other data sets.

To compare the output and results of the rolling regression model of WTI returns and the sentiment variable, an additional two other rolling regression models are computed by the system with two other key variables included in each model. One additional rolling regression model incorporates news volume from the FT crude oil corpus. The news volume from this corpus fluctuates from day-to-day and may contain explanatory information for WTI returns, this hypothesis is evaluated by the system in this section. Another rolling regression model tests the hypothesis that WTI trading volume has explanatory power for WTI returns. Trading volume is chosen here as typically it is used as a proxy for market sentiment and acts as a comparison for the negative sentiment variable to see if any explanatory information is contained in trading volume.

Three rolling regression models are computed to assess differences in statistical power of negative sentiment, WTI trading volume, and the news volume of the FT crude oil corpus. The results in this section show three figures representing the statistical significance of the hypothesis tests for the inclusion of three variables. Table 4.22 shows the number of statistically significant hypothesis tests as a percentage of all models estimated for the rolling regressions.

Table 4.22: Rolling VAR (250 day window) model specified in Equation 4.4 for the period 2001-01-10 to 2014-12-08 (n = 3303). The significance is measured as the hypothesis test for the inclusion of the exogenous variable. The table shows the VAR estimation with columns as the independent variable and rows the dependent variable. The result shows the number of significant models as a percentage of total models estimated for the time period.

	Statistically Significant models			
	Returns	Negative Sentiment	News Volume	Volume Traded
Returns	-	23%	21%	3%
Negative Sentiment	13%	-	10%	8%
News Volume	17%	13%	-	22%
Volume Traded	27%	16%	24%	-

The results of the rolling regression show differences in the explanatory power of the sentiment variable, news volume, and trading volume over time. The negative sentiment variable is not always statistically significant in predicting returns. The sentiment proxy is seen to reject the null hypothesis and contribute as a statistically significant explanatory variable 13% of time as seen from the results in Table 4.22. News volume more frequently impacts returns and a strong relationship is seen between news volume and trading volume. News volume is seen to impact trading volume, which in turn has the highest number of significant models for predicting returns. Returns predict negative sentiment more than news volume and perform poorly when predicting trading volume. The time series plots were computed by the system as described before (Section 4.2.4), where a value of one represents a hypothesis test that was significant at the 90% level and a value of two for 95% and 99%. The results are displayed in Figure 4.6, Figure 4.7, and Figure 4.8.

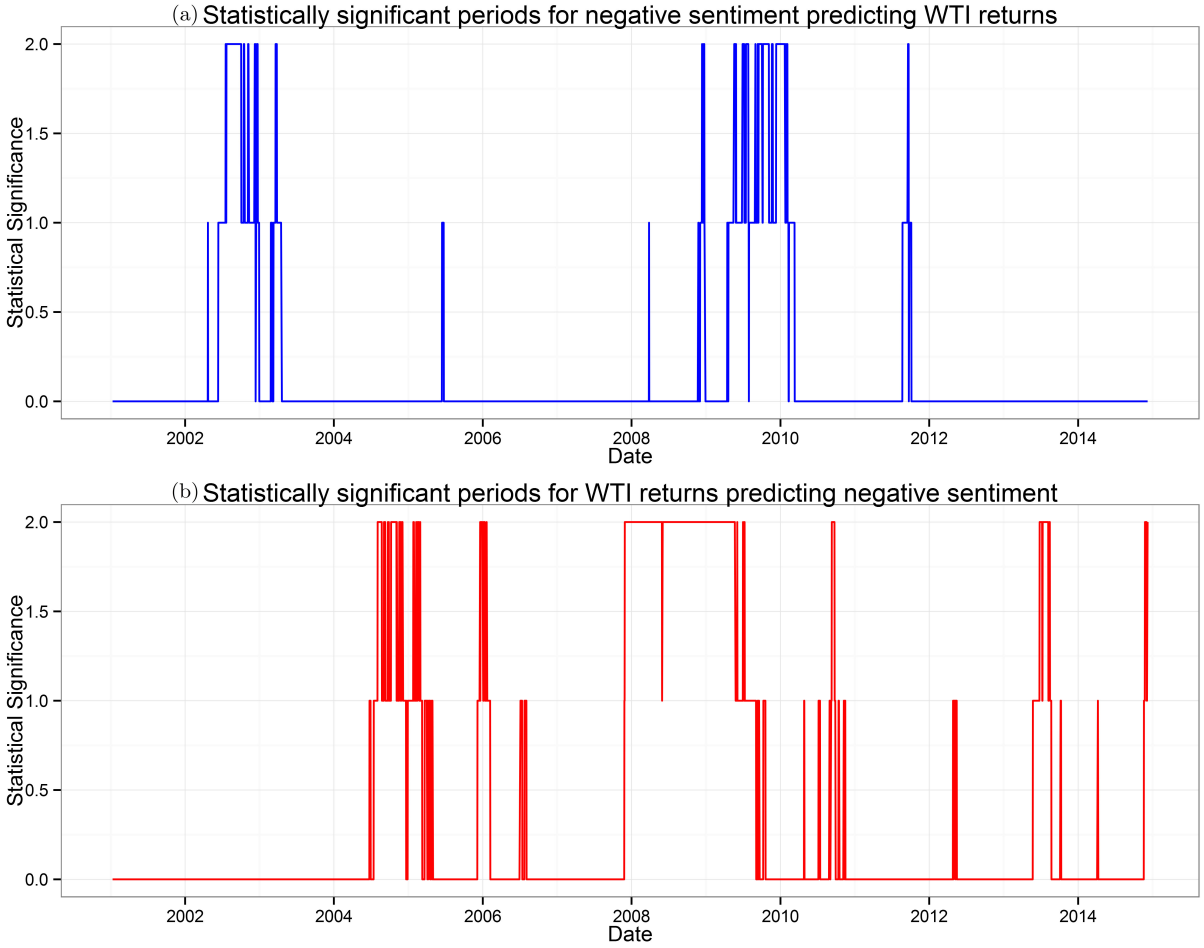

Figure 4.6: Rolling VAR (250 day window) model with WTI futures returns and negative sentiment for the period 2001-01-19 to 2014-10-31 (n = 3267). Negative sentiment predicts WTI returns (a) and returns predicting negative sentiment (b). The negative sentiment proxy was derived from the FT crude oil corpus. The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

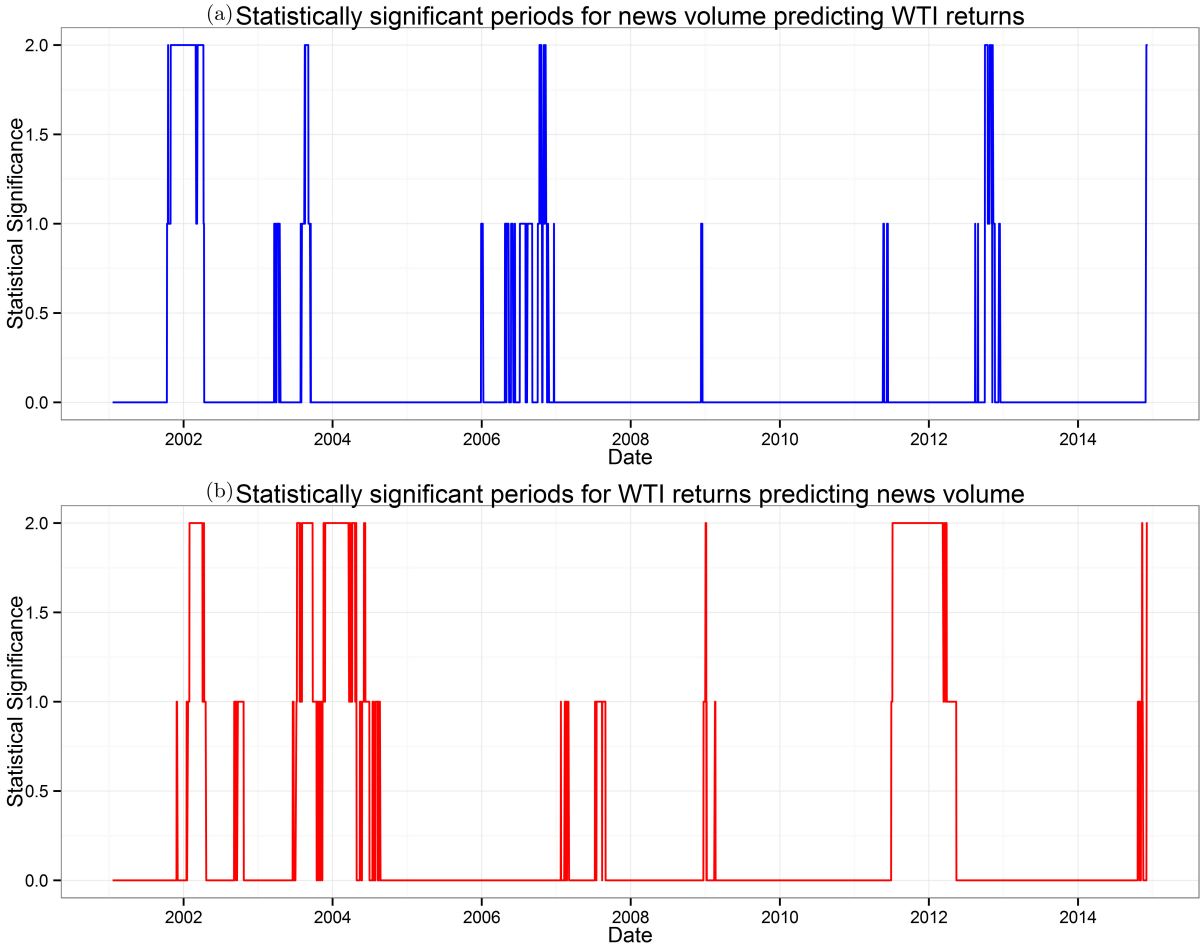

Figure 4.7: Rolling VAR (250 day window) model with WTI futures returns and news volume for the period 2001-01-19 to 2014-10-31 (n = 3267). Negative sentiment predicts WTI returns (a) and returns predicting negative sentiment (b). The news volume was derived from the FT crude oil corpus. The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

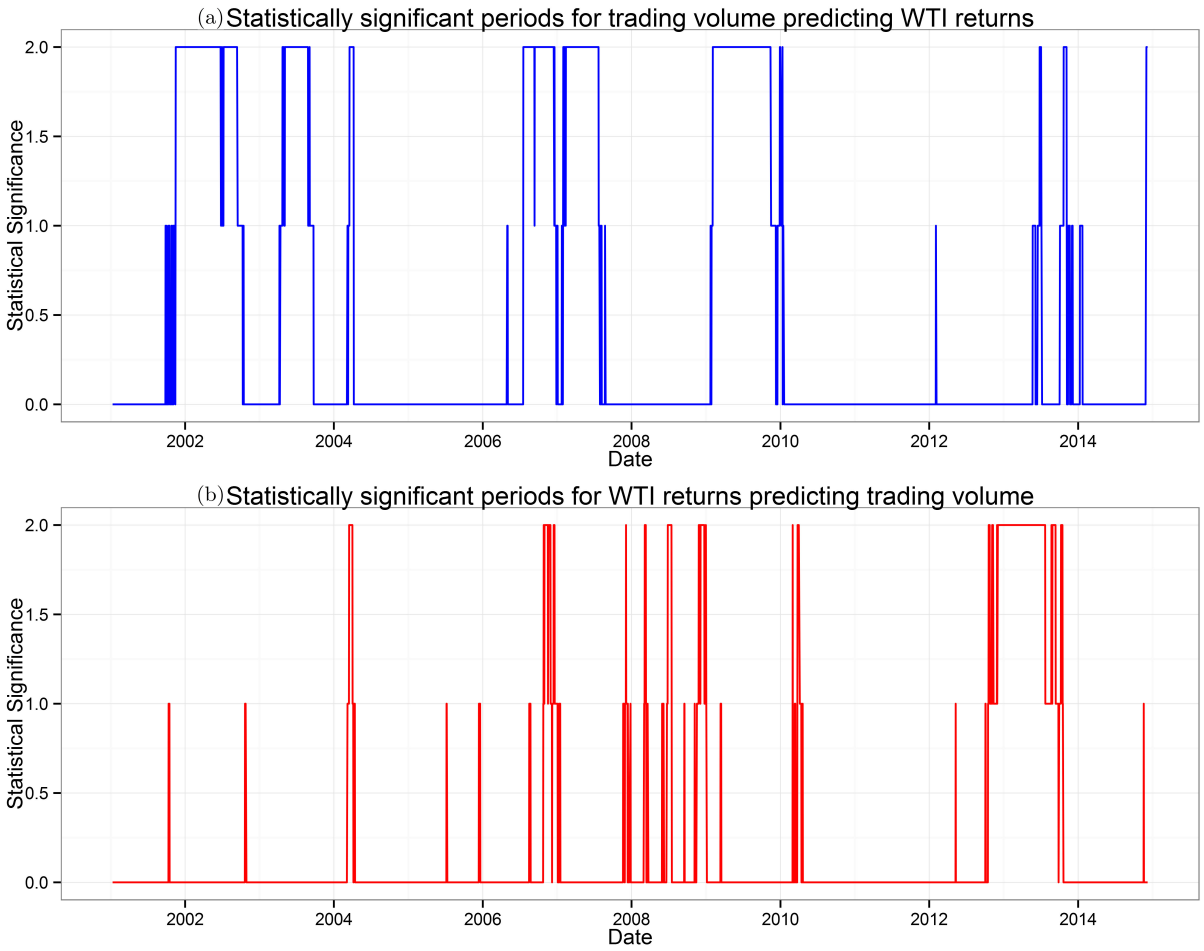

Figure 4.8: Rolling VAR (250 day window) model with WTI futures returns and WTI trading volume for the period 2001-01-19 to 2014-10-31 (n = 3267). Negative sentiment predicts WTI returns (a) and returns predicting negative sentiment (b). The significance is measured as the hypothesis test for the inclusion of the exogenous variable. A value of one shows significance at the 90% level and a value of two for 95% and 99%.

The results reiterate the changing influence of sentiment on returns in time seen in the equity market case study in Section 4.2.4. The average impact of negative sentiment is shown to have the greatest statistical significance and robustness while the rolling regression results suggest the predictive power of sentiment, news volume, and trading volume changes in time. The use of the rolling regression models highlights the changing impact of sentiment, whether this is due to market enthusiasm, volatility, or other anomalies and events remains to be investigated fully. This section has demonstrated a similar evaluation of the system and the time varying modelling component using different data with a result being computed that corroborates the results of the previous evaluation (Section 4.2.4).

4.3.4.1 Business Cycles and Sentiment in Commodity Markets

This section evaluates the results of computing sentiment with the business cycle model described in Equation 4.2 for the commodities market. Results in the previous case study used the system to compute a model that showed a link between negative sentiment and the changing business cycles. To examine this link for the commodity market a model is estimated as before where the returns (r_t) series is for WTI future returns. The sample period is from 2000 to 2014 incorporating negative sentiment from the FT crude oil corpus.

Equation 4.2 contains two dummy variables for recessionary and expansionary periods. Two recessionary periods are recognised during the period of the sampled data used in this evaluation. The first recessionary period begins in April 2001 continuing to November 2001 and the second period begins on January 2008 and lasting until June 2009.

Table 4.23: Coefficients for the negative sentiment variable computed from Equation 4.2. Negative sentiment was computed from the FT crude oil corpus for the period of 2000-01-04 to 2014-12-08 (n = 3553). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	Recession	Expansion
$NegSent_{t-1}$	2.8	-4.0
$NegSent_{t-2}$	*-11.9*	-4.7
$NegSent_{t-3}$	**-17.3**	-2.4
$NegSent_{t-4}$	-0.2	*-6.3*
$NegSent_{t-5}$	6.0	-1.1
$\chi^2[NegSent]$	***30.37***	7.25
AIC	-16505	-16505
$\bar{R}^2$	1.2%	1.2%

The coefficients computed for the negative sentiment variable during recessionary periods and expansionary periods are presented in Table 4.23. An increase in the impact of negative sentiment is noted during recessionary periods than in expansionary periods. Although different from the effect seen for the equity market case study, this result corroborates the study by Garcia [42], where a media proxy (ratio of negative and positive sentiment) is found to predict 12 basis points in recessionary periods versus 3 basis points in expansionary periods for DJIA returns. The impact of sentiment is greater in recessions than during periods of expansion. The effect of negativity is dispersed across the lags of the sentiment variable. The negative sentiment variable has a greater impact on WTI returns with no evidence of a reversal over the five lags.

The hypothesis test shows the significance for the inclusion of the sentiment variable during the recessionary period (30.37 at 99% confidence) but does not demonstrate a high degree of confidence for its explanatory power during expansions (7.25 with less than 90% confidence). The results presented in Table 4.23 attests to the conjecture that news content is more readily absorbed in downward business cycles than in more exuberant times.

4.3.4.2 Summary

The result of using a moving window and the determination of different business cycles indicates that the effect of the sentiment proxy varies through time for the commodities market. Financial markets are seen to be irrational and rational, deviating from fundamentals and adhering to modern financial principles at different times. Due to this, the significance of some variables may change in time, methods of detecting when these changes will appear and assessing the impact of variables will greatly improve the explanatory power of models of financial returns. By using a moving window method it is possible to avoid some time series anomalies and certain market conditions. By observing the results of different models in time as computed by the system, results have shown that business cycles and the state of the market as a whole can change the impact that the sentiment variable can have on returns. The business cycles data was incorporated from the NBER and it was found that negative sentiment is more heavily correlated with recessionary periods than expansionary periods. The results computed by the system for the changing influence of the sentiment variable on returns in time are similar and comparable to those presented in the previous case study.

4.3.5 Commodity Market Volatility and Sentiment

The behaviour of investors to sentiment may induce market volatility in the commodity markets the same as in other financial markets. The evaluation of the sentiment variable and equity market volatility was estimated by the system and results presented in Section 4.2.5 and showed tentative links between sentiment and equity market volatility (Section 4.2.5.1). In this section the relationship between sentiment and commodity market volatility is examined following the same method and estimating the same model as defined by Equation 4.3 using WTI returns and the negative sentiment variable from the FT crude oil corpus.

WTI returns are used to calculate volatility, negative sentiment and the daily news volume computed from the FT crude oil corpus, and WTI trading volume are incorporated as independent variables into the regression model. The volatility series is computed by the system as the log of the conditional standard deviation calculated using a GARCH(1,1) (v_t) model. The independent variables in the regression include the log detrended trading volume of WTI returns (vlm_t), the negative sentiment variable derived from the FT crude oil corpus (s_t), and news flow which is the number of articles published per day in the FT crude oil corpus ($svlm_t$). All variables are stationary at the 99% level according to the Augmented Dickey Fuller (ADF)

test (Appendix B, Section B.1, Table B.2). The transformations and model have been described are described in more detail in Section 4.2.5.1.

Table 4.24: Coefficients estimated for Equation 4.3 for the period 2000-01-06 to 2014-11-19 (n = 3535). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	WTI
v_{t-1}	***3303.5***
vlm_{t-1}	***47.0***
s_{t-1}	11.4
$svlm_{t-1}$	11.5
AIC	-12202
$\bar{R}^2$	98%

Table 4.24) shows the results computed by the system for estimating Equation 4.3 with WTI returns. WTI trading volume and previous day volatility are seen to have a significant impact. As the volatility estimation relies on the conditional values of variance, a large significant impact is expected with a large adjusted r-squared value ($\bar{R}^2$) as much of the variance is explained or already accounted for with this term. No statistically reliable result is seen for news volume or negative sentiment on WTI returns. Despite this news volume is again shown to predict next day volatility where an increase in the volume of news will result in an increase in volatility. Results are smaller in magnitude when compared to the standardised results of previous day volatility and trading volume.

The system computes a similar magnitude for the impact of news volume and sentiment on WTI volatility. However the statistical confidence for each of the coefficient values is low in each case. News volume is noted as having a larger impact on volatility than sentiment alone as reported in Antweiler and Frank [8]. Although one news type was studied in this case, the volatility regression for WTI may benefit from further investigation into the type of news used to create the news proxy.

4.3.6 Summary

The preceding section has evaluated the implementation of the system described in Chapter 3 by using financial data from the commodity market and domain text collected from the *Financial Times* and the *Oildrum* blogs. Changes in WTI price is used to evaluate the influence and relationship of the sentiment variable computed on commodities. If is found that negative

sentiment also plays a role in commodity markets, with the negative sentiment variable predicting up to an -8.5 and -9.5 basis point change in WTI futures and spot returns respectively. A cumulative impact of -16.4 and -17.6 basis points is seen across all lags of the sentiment variable that is only marginally reversed across the five lags. An evaluation of the content analysis method has been presented with the differences in computing the sentiment variable from formal and informal media compared. The results of using different text types and genres to create a proxy were shown, with consideration given to improving a general language dictionary with domain terminology. The time varying impact of negative sentiment on WTI returns was examined with similar results to the previous equity market case study being found. Sentiment is seen to only sometimes predict returns as computed using the rolling regression model from the system's modelling component. Results incorporating business cycle effects show negative sentiment to have a much stronger impact on WTI returns in recessionary periods (up to -17.3 basis points) than expansionary periods corroborating results from the literature of news absorption and market timing. Lastly, conclusions about the interaction between sentiment and volatility are drawn.

4.4 Conclusion

In this chapter an evaluation of the system described in Chapter 3 as been presented and illustrates how the content analysis and modelling components of the system have combined methods from two disciplines to compute a sentiment variable that has statistically significant explanatory power for determining changes in financial assets. As compared to previous studies, the system removes much of the burden on data collection, constructing a time ordered corpus of text, performing text analysis and computing a time series of sentiment, aggregating and aligning this output with automatically retrieved financial time series data, and finally estimating a statistical model to determine inter-relationships between variables. This has been illustrated by the case studies presented here using different text and time series data for two different financial markets. Using the system the differences that the choice of text data, choice of financial instrument, and choice of model have on results has been investigated and evaluated throughout this chapter.

For the text analysis component of the system, the choice of news and text plays a vital role in determining the statistical confidence of the sentiment variable computed in both financial markets. Agenda-setting columns such as the *Abreast of the market* in the WSJ and the *Lex column* in the FT are a good source of information that can reflect and influence investor sentiment. Sources that are respected, having a good reputation with investors, high readership, are well established and can be useful when it comes to creating a proxy for investor sentiment. Market specific news also plays a more substantial role in explaining movements for financial instruments. Industry specific news can act help create a better proxy for predicting changes in instruments in a particular market. News about oil production, policy, and the strength of the

industry acts as a better proxy for market or investor sentiment specific to an asset class such as commodities for instance. Differences are also drawn between informal and formal news. While both have predictive power for asset returns, the difference lies in the degree of explanatory power. Overall the importance in the choice of news in creating a valid proxy for news is evident as evaluated by the different text corpora used by the system.

The algorithm used by the content analysis component showed some differences when including a domain lexical resource with that of a general language dictionary. A glossary of oil industry words from a reputable source Platts and Oil and Gas UK was seen to aid the content analysis capabilities of the system. Words and phrases in the domain glossary typically carry no sentiment bearing terms, referring mainly to industry specific words. The use of domain specific terminology provided an easily adaptable solution for the dictionary based approach. A reduction in the frequency of negativity was seen when extracting sentiment from domain text. This marginally affected the impact of the negative sentiment variable on WTI futures returns. The predictive impact of the sentiment variable from informal domain news was seen to increase while little change was noted for formal media.

In the models computed by the systems, measures of negative sentiment had a stronger relationship with DJIA and WTI futures returns than positive sentiment measures from the same text corpora. Theories from behavioural finance suggest that the effect of sentiment is asymmetric [54]. Market pessimism seen in investor behaviour is also believed to add doubts to the abilities of rational investors to arbitrage and thereby correct market inefficiencies [18]. Studies involving textual sentiment have corroborated the higher correlation of returns with negative words [96] [40]. The results of each case study show an asymmetric response of news on the returns of each asset class, where negative sentiment is a better and more statistically significant explanatory variable for returns. This is in support of the original behavioural model proposed of noise traders and irrational behaviour in financial markets [31].

Differences in the impact of sentiment on returns over time are noted for equity and commodity markets. From computing the rolling VAR models and hypothesis test of sentiment and returns in both case studies, returns are seen to have a higher frequency of predicting sentiment than sentiment has on predicting returns. Large shocks in returns and prices in markets will usually impact investor sentiment and are often then reported in news, while shocks in sentiment will only sometimes impact returns and prices, even if covered in the news. The system found the statistical confidence of the sentiment variable to vary in time for explaining changes in returns in both markets. A VAR model that incorporated business cycle information also showed the explanatory power of the sentiment variable to change in time. A greater impact and absorption of news during economic downturns was seen in the case of the commodities market, corroborating previous studies [42]. The system demonstrated that the interaction between the DJIA and commodities market and the content of news media changes through time. Sentiment impacts returns periodically, while returns impact sentiment much more frequently. The parameters of returns distributions change through time as does the impact of sentiment.

The capability of the system to be updated each day with new data scrapped from the various data sources also showed the benefit of a computational systems approach.

The impact of information and news has been considered intertwined with volatility in financial markets. With news arrivals clustering in the same way volatility does. Increased variance in a market has been interpreted as new or difficult to process information. This would indicate that news would closely correlate to changes in variance. The DJIA series is shown to have higher volatility than the WTI series, as evident from the results of the baseline GARCH(1,1) model. A number of explanations can be cited for this, one in particular is the regulation of oil markets. One would expect sentiment to have less impact in the oil markets than the equities. Tight restrictions and control, regulation, and the fact that commodities are tied to a physical quantity may mean they more closely reflect fundamentals. Equities are more favoured by the public and as such may be more susceptible to speculation. Although evidence has been given to this conjecture, the text derived proxy presented here shows a marginal relationship to typical volatility estimation and demonstrates a promising avenue for future examination.

The information extracted from news and text by the system whether it acts as a proxy for investor sentiment, market sentiment, noise trader activity, differences in transaction costs, or liquidity effects, has been demonstrated to have explanatory power for returns. Rarely would this information derived from news and text be considered new information but instead would be information that is already known. The type of news and text supplied to the system influences the explanatory power of the sentiment variable produced. Opinion columns are seen to predict an initial response that is quickly reversed in subsequent returns with no lasting or long term effects being observed. Typically, news will be about future events, opinions of these events, and the implications for the business industry, response to this information has been said to be that of noise traders, where the proxy extracted from text is a representation of investor sentiment.

The system implementation evaluated in this chapter provides a platform on which to process text corpora of different topics and subject matter while using the same algorithm and can be augmented with different dictionaries or word lists as inputs. The output from the content analysis component can be merged and aligned with any financial time series data, which can also be retrieved from a number of database and online sources. Several statistical models can be estimated to evaluate the impact and statistical significance of the sentiment variable on changes of a financial instrument. The work in this chapter has provided a thorough evaluation of the system implemented and described in Chapter 3. The system demonstrates a number of automated tasks as a result, including text and time series collection, data processing and aggregation, model estimation, and visualisation. The system removes these burdens from an investigator allowing a result to be computed for the influence of a news source on any financial asset outputting a result and interpretation that can be processed by an investigator more easily and efficiently.

5 Conclusion

The influence that news media and announcements have on financial markets is well known. Having been thoroughly investigated in the literature, attempts to incorporate their effects have been numerous and vary widely. In more recent times, methods drawn from content analysis and machine learning have become popular as a way to measure the content of news in finance. It has been proposed that the tone of news can be estimated by counting the number of *affect* terms that occur in an article. The tone estimated from news has been shown to predict financial returns. The predictability of news on financial markets has been mostly interpreted in two ways in financial literature; news reports on information that has not been fully incorporated into price and markets, or news influences the attitudes and beliefs of investors prompting a response. Studies in finance that have studied the content and tone of text have ranged from news articles and social media to earnings releases and corporate disclosures. The literature and research studies using news and textual sentiment have been growing steadily. The application of text analysis methods to financial modelling has been diverse. The work presented in this thesis has also contributed to this growing area of study.

The main objective of this thesis has been to develop and evaluate a system that can be used to compute a time series of sentiment from a corpus of news and to use this variable in a statistical time series model to the returns of a financial instrument. The system has successfully integrated a number of functions to achieve this objective. For the system implementation it was necessary to have functions that could harvest and scrape data from online sources, perform data processing and aggregation, content analysis, and time series modelling. The framework and implementation of the system can take different text sources and financial time series data

to be input to the system allowing it to analyse the impact of news on any financial market. The system has been evaluated on data from two financial markets with a number of different text types and using different time series models.

The system has been used to evaluate the extent that sentiment as a variable is a contributing factor for asset returns. This was done by examining the statistical significance and explanatory power that news sentiment extracted from the content of news has on asset returns. By looking first at the equity market the predictive impact of a proxy for sentiment from text is verified. This response is then evaluated on another market, that of commodities, showing that a proxy for sentiment extracted from news and text has explanatory power for more than one type of asset.

Central to the case studies and evaluation that have been presented is the construction of a text corpus that relies on choosing a news source, type of article, and topic. The impact that different text types can have on the results and explanatory power of the news media proxies was examined in depth. The consistency of news, its topic of discussion, type, and source play a role in the performance of the sentiment proxy extracted. News that is on topic and relevant or specific to a domain, such as the opinion columns *Abreast of the Market* from the WSJ and the *Lex* column from FT which both discuss market level news, or the *Oil & Gas* section of FT which discusses oil industry news, have a more consistent impact on returns than general widely sampled news. The type of news is shown to be paramount in creating a corpus and text based proxy in modelling affect and assessing the influence of news.

The time varying effect of news and text sentiment on returns was also investigated. The level of sentiment about assets and markets has been known to change in time and has been studied previously, incorporating content analysis methods to summarise text sentiment [42]. Evidence for the changing impact of sentiment across asset classes was found and links made with the impact of sentiment according to different business cycles.

Lastly, sentiment and news volume were seen to relate to volatility in different asset classes. Tentative links were made between news content, in the form of the negative sentiment proxy, and the volume of news in explaining raw volatility for each of the assets.

The remainder of this chapter will discuss the contributions of this thesis and the context of these results in an inter-disciplinary domain of content and sentiment analysis in finance (Section 5.1). A discussion of the limitations of the work is the presented (Section 5.2), followed by contributions and potential future work (Section 5.3).

5.1 Contributions

There exists a wide range of approaches to modelling financial time series and predicting changes in financial markets. Methods for text analysis and analysing news text are also numerous. A large body of literature exists that has attempted to combine methods from each discipline. Chapter 2 gives an overview of content analysis methods as applied to finance and financial

modelling. Many of the studies have focused on a particular news type or market both of which are chosen subjectively. Some studies describe a framework or system that is catered towards analysing a particular source of news or text for a financial instrument.

The main contribution of this thesis is in the development of a system that can compute a sentiment time series from a collection of news articles or time annotated text. This sentiment variable can then be aggregated with a financial time series in a statistical model to examine any inter-relationship between news and financial markets. The system framework described in Chapter 3 allows any text corpus to be input and and time series to be imported as well. The domain independent approach allows different data sets to be used with the system. The system allows different news types to be used to estimate the sentiment variable and the influence of news in different financial markets. This allows a less subjective analysis to be carried but also facilitates an investigation into the relationship between different text types and time series data. The system and underlying framework allows a number of functions to be carried out to perform this task such as:

Collect and scrape text and financial data from different sources, constructing a corpus in XML for text data and retrieving time series data.

Compute a sentiment time series using a component that performs content analysis by importing the structured text corpus and a general language and domain dictionary or word list.

Estimate time series models and hypothesis tests to investigate the explanatory power of the sentiment variable with the financial time series data.

Output and store the results of the estimations and visualise the results.

The system is evaluated using different data inputs and estimates the impact that different news sources and types have on changes in the price of assets for two financial markets. This evaluation takes the form of two case studies in Chapter 4. By using the system, human intervention is kept to a minimum. The burden of data collection, analysis, and interpretation are removed from a user reducing the chance of error in estimation.

The case studies presented in Chapter 4 evaluate the implementation for the system and methods used to construct the sentiment proxy from news text. To examine any inter-relationships between financial markets and the sentiment variable, a number of time series models are estimated in the modelling component of the system to examine the impact of news on returns. A VAR specification is estimated with control variables that account for market anomalies and seasonality. These control variables account for past market returns, the effects of trading volume, volatility, and the weekend and January effects. The control variables are an attempt to isolate the effect of sentiment by including variables that may also account for the same information. The impact of sentiment is seen to be consistent as the VAR model is estimated by the system incrementally including the control variables one at a time to see if any confounding effects with the sentiment variable are seen. In the case of the DJIA, results show a -4.7 basis point impact on DJIA returns from a one standard deviation change (increase) in negative sentiment from

the WSJ *Abreast of the Market* column. General news from the WSJ does not have the same predictive power suggesting that the choice of news is an important factor in constructing the news proxy. This result is robust, remaining statistically significant after extending the data set in time.

To check the influence that news type and source have on this result another agenda-setting column, the FT *Lex* column is used to generate the negative sentiment proxy. It is seen that negative sentiment extracted from the *Lex* column has a larger but similar negative impact on DJIA returns at -8.5 basis points. The results show the average impact of sentiment on DJIA returns and the ability of the system to compute a sentiment variable that is statistically significant and has explanatory information for changes in financial returns. A rolling regression model is also estimated by the modelling component of the system, which is used to assess the explanatory power of sentiment in time. Applying this method to DJIA returns and using the negative sentiment proxy as an independent variable, it is found that returns predict sentiment on more occasions than sentiment predicts returns. News will often report and reflect on events that have occurred and influence markets. The information will have already have been incorporated into price while publications report on events retrospectively. Using the rolling regression model, sentiment is seen to benefit the explanatory power of the model for returns some 21% of the time. The average impact of negative sentiment is associated more with expansionary periods for DJIA returns than recessionary times but its magnitude in expansionary periods (-3.81 basis points) is similar to the average effect over the entire period (-4.7 basis points) without accounting for business cycle effects.

The system uses data from a different financial market that of commodities along with domain news to further investigate the results computed in the first case study. The oil market draws a lot of attention from media where news ranging from geo-political events to natural disasters are reported and their subsequent effects on oil commented upon. Oil is a fundamental product that many industries are dependent on and influences the world economy. With high levels of regulation and frequent announcements, the oil commodity market is an information rich market. The algorithm implemented in the content analysis component of the system aids in the computation of the sentiment proxy by incorporating domain terms and phrases. The sentiment variable computed is used in the modelling component to estimate the influence that a proxy for domain news has on West Texas Intermediate crude oil. The average impact of negative sentiment extracted from oil related news in the Financial Times shows the proxy can account for up to a -8.5 basis point change in WTI futures returns and up to -9.4 basis points for WTI spot returns. The impact of the negative sentiment proxy is dispersed across the five lags of the sentiment variable, with a cumulative (statistically significant) influence of -16.4 and -17.6 basis points for WTI futures and spot returns respectively.

The influence that different text types has on computing the sentiment variable is evaluated by using different text corpora as an input to the system. These text corpora include oil specific news and oil blogs. Using these corpora the difference that formal and informal news has on the

sentiment variables construction can be investigated. By using three different text collections, the change in the resulting coefficients and model incorporating the sentiment variable with returns for each corpus was evaluated. The negative category of the General Inquirer dictionary was shown to perform adequately for constructing a proxy from formal news; the Financial Times crude oil corpus. Filtering this word list with a domain glossary was shown to create a proxy that predicted a higher impact from negative sentiment for domain related informal news; the Oildrum blogs. This latter investigation formed the basis of evaluating the content analysis component and algorithm described in Chapter 3 Section 3.3.2.

The time varying influence of the sentiment variable produced by the system with the commodities financial data was also evaluated. The influence and statistical significance of sentiment was seen to again vary in time, with negative sentiment only sometimes predicting WTI future returns. The results incorporating the business cycles show that negative sentiment has a stronger impact in commodity markets during recessionary periods (up to -17.3 basis points) than expansionary periods. This result is more in line with previous studies by Garcia [42] who suggest investor sentiment and the reaction of market participants to news is greater during economic recessions than expansions.

The system utilises open source tools and freely available datasets. Financial data in particular has seen an increase in availability in recent times with vendors such as *Quandl* adopting new business models to freely supply high quality data with limited restrictions. The increased data availability, the continued growth of the internet, and pervasiveness of open source technology has meant the volume of data being created and released is increasingly rapidly. With new tools and frameworks being proposed, applications and systems such as the one developed here that utilises a multitude of resources and data sources to perform novel data analysis will benefit greatly.

5.2 Limitations

One limitation of the work presented here relates to whether news text is indicative of real information for financial markets. This brings into question the content analysis based approach of extracting a meaningful measure of sentiment from text. The assumption of the method is that affect and the tone of a text can be summarised by counting the frequency that special pretagged terms from a lexicon appear in a text. In the work presented in this thesis it is assumed that a meaningful and well-structured dictionary is used in the content analysis method, this would include the GI dictionary and terminology work done by Stone et al. [92]. Whether these terms are still interpreted as their initial interpretation or their use has changed through time is a consideration for the validity of the GI program and of content analysis methods using this dictionary. Machine learning approaches allow a new model to be trained and updated as new text becomes available. By choosing a representative tagged corpus, a model can learn the most up to date usage of terms and phrases. This approach however, is not as linguistically

motivated. The choice of negativity is no longer in the hands of experts but instead at the choice of the experimenter or defined by a tagged sample. Neither approach is completely objective or convenient but this does not detract from the usefulness of the application.

The application of content analysis methods to the domain of finance adds another assumption that this is a useful approach to create a new explanatory variable that can be used in financial and time series models. Those reading the news are influenced by the content and sentiment contained within the text and that their perceptions and actions are influenced as a result, which is then reflected in market movements. Once the text has been analysed, the output is combined with a quantitative model to examine the relationship of market participants to news. Regression analysis is one the most widely used method in financial literature and time series analysis with analysts identifying relationships among variables and testing hypothesises. The goal of this analysis is to identify which variables have an impact on the dependent variable and to assess the strength of the relationship. The evaluation in Chapter 4 looked at how strong and statistically confident the independent variables effects were on the dependent variable, which was reflected in the t-statistics and generalised with the p-values of the computed coefficients. The relationship, described as the average effect, looked at a unit increase in the independent variables implying the coefficient value increase in the dependent variable on average. This is an acceptable method for revealing the importance and relevance of independent variables of interest [91]. This allows the results to be represented and the importance of the findings to be expressed. A relationship between variables can be characterised using regression. However, prediction on the back of these results can reveal the limits of such models. Variability exists around the effects and results determined. This needs to be communicated in the reported results. The system architecture and evaluation described in Chapter 3 addressed this issue somewhat by allowing for different models to be used interchangeably. The modelling relies on describing explanatory links between the proxies for news media and changes in returns. VAR, hypothesis tests, and rolling window methods are all employed by the modelling component to assess the statistical significance and the explanatory power of the sentiment variable in a number of ways. Less emphasis is put on forecasting future values of the dependent variable. Nevertheless, the literature gives evidence that linear models are suitable and a common method to examining potential effects between variables. Much of the recent and popular literature that forms the basis of this study has utilised such models. Using parsimonious models and verifying results in a robust manner creates a firm foundation to then explore using different models to examine the dynamics in the data. The development of the system allows different models to be quickly substituted and can form the basis of future work and evaluation.

5.3 Future Work

Having described the contributions of this thesis, potential future work is now presented. The contributions of this thesis include a system that has successfully been evaluated and shown that

news text has a measurable influence on financial returns. Evidence is given for the time-varying effect of the sentiment variable on returns also. Although additional data (information about the business cycle) and statistical methods (rolling regression) are used in conjunction with the regression methods to show this varying effect, further investigation would be necessary into answering why sentiment predictability changes in certain periods. Some evidence is given to support the idea that information is absorbed differently according to business cycles. The state of economy plays a role in the response of investor sentiment and influence of news media. The content of news and period of reporting may also be linked to the response seen in the news proxy. Additional studies focusing on the relationship between the response of financial markets and sentiment from text may help to understand the complexity of this relationship.

Using time series models that can distinguish between the short-term and long-term effects of sentiment computed from news text would be an interesting avenue of research. The use of structural VAR to investigate the contributions presented here, in particular looking at the increased amount of variance being explained with the news media proxies, would help in corroborating the relationship suggested between variables and the short and long-term impact of sentiment on financial returns.

Investigating additional models and their performance would address issues of model uncertainty. The system's components and framework allow a forecasting model to be incorporated. This implementation could draw more on machine learning techniques to quickly assess the performance of the sentiment variable and how it varies in different estimations. With the increasing availability of tools, applications, architectures, and software libraries, a multitude of models and their performance can be evaluated with relative ease. As larger volumes of new and diverse data become available ways of addressing model uncertainty will be needed [103]. The framework of the system described in Chapter 3 would allow the issue of model uncertainty to be more easily addressed than would otherwise have been achieved before.

5.4 Summary

This chapter has summarised the main body of work in this thesis, potential limitations and shortcomings of the approach, and avenues for future research. The system developed and described in this work combines a number of methods from text analysis, financial time series analysis and econometrics. The text analysis component is used to collect, aggregate and compute sentiment from the content of news and text from different sources and of varying text type. The extracted sentiment variable has been linked to the movement of a market index (DJIA) and a benchmark for the commodities market (WTI). This has shown the flexibility of the system and its ability to estimate the relationship between sentiment and financial returns using different data and models. The results corroborate the literature; negative sentiment extracted from an agenda setting column impacts returns, negative sentiment is more statistically significant and has a greater impact than positive sentiment suggesting an asymmetric impact

for market indices and commodities from news media. The time varying impact of sentiment and relation to volatility highlight the complexity and depth of application of textual sentiment in financial markets. The future work also presents areas of potential investigation. The usefulness of a proxy for information extracted from text and news media using a computational approach is apparent, with a multitude of new avenues of research and applications yet to be investigated.

Computers are involved in most economic transactions and account for the majority of trades executed in financial markets. With the deluge of information set to grow and come from new and unconventional sources, techniques and tools for manipulating this data, making sense of it, and acting upon it will become essential. As new problems arise with the quantity of data and methods of making sense of it, so too will the opportunities. It is the intent that the work presented in this thesis has contributed to making sense of this information in new and unconventional ways and that the techniques, methods, and system presented will aid in understanding the role of textual sentiment in finance.

A

Appendix: Case Studies - Equities and Sentiment

A.1 Diagnostic Tests

Table A.1: Results for Augmented DickeyFuller test for stationarity up to five lags. DJIA, Negative sentiment, volume, and volatility relate to the time series used in the market sentiment section. Abreast of the market uses time series from 1989-01-03 to 2008-08-25 (n = 4789). The Lex columns is for the period 2005-01-030 to 07-11-2014 (n = 2409) the VIX index is used as the volatility measure. All time series are tested using short (5-lags) and long(order determined by AIC) lag order

	ADF with 5 Lags	
	Aotm	Lex
DJIA	***-30.41***	***-21.43***
Negative	***-22.45***	***-15.52***
Volume	***-24.4***	***-16.4***
Volatility	***-26.1***	***-23.77***

Table A.2: Results for Augmented DickeyFuller test for stationarity up to five lags. Volatility v_t is a log of the conditional standard deviation from a GARCH(1,1) for DJIA returns. Volume vlm_t is the trading volume of each instrument, s_t is the sentiment proxy from the Lex column for the DJIA column, news flow $svlm_t$ is the number of articles per day. The period includes 2005-01-05 to 2014-11-07 (n = 2411). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	ADF with 5 Lags
	DJIA returns
v_t	***-21.43***
vlm_t	***-16.40***
s_t	***-15.52***
$svlm_t$	***-6.96***

B

Appendix: Case Studies - Commodities and Sentiment

B.1 Diagnostic Tests

Table B.1: Results for Augmented DickeyFuller test for stationarity up to five lags. WTI futures, Negative sentiment, volume, and volatility relate to the time series used in the commodity sentiment section. All time series are tested using short (5-lags) and long(order determined by AIC) lag order

	ADF with 5 Lags
	FT oil corpus
WTI futures	***-25.525***
Negative	***-18.399***
Volume	***-25.09***
Vix	***-28.196***

Table B.2: Results for Augmented DickeyFuller test for stationarity up to five lags. Volatility v_t is a log of the conditional standard deviation from a GARCH(1,1) for WTI futures. Volume vlm_t is the trading volume of each instrument, s_t is the sentiment variable from the FT crude oil corpus, news flow $svlm_t$ is the number of articles per day. The period if for 2000-01-06 to 2014-11-19 consisting of 3535 observations. Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 1% (Bold Italic), 5% (Bold) and 10% (Italic) levels. Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals.

	ADF with 5 Lags
	WTI returns
v	***-4.4266***
vlm	***-24.886***
s_t	***-18.578***
$svlm$	***-13.334***

Bibliography

[1] K. Ahmad. Being in text and text in being: Notes on representative texts. *Incorporating corpora. Clevedon: Multilingual Matters*, pages 60–91, 2008.

[2] K. Ahmad, C. Kearney, and S. Liu. No news is good news: A time-varying story of how firm-specific textual sentiment drives firm-level performance. In *The European Financial Management Association 2013 Conference*, 2013.

[3] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[4] T. G. Andersen. Return volatility and trading volume: An information flow interpretation of stochastic volatility. *The Journal of Finance*, 51(1):169–204, 1996.

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

[6] T. G. Andersen, T. Bollerslev, F. X. Diebold, and H. Ebens. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76, 2001.

[7] A. Ang, G. Bekaert, and M. Wei. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of monetary Economics*, 54(4):1163–1212, 2007.

[8] W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

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

[10] S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[11] F. Black. Noise. *The journal of finance*, 41(3):529–543, 1986.

[12] B. J. Blair, S.-H. Poon, and S. J. Taylor. Modelling s&p 100 volatility: The information content of stock returns. *Journal of banking & finance*, 25(9):1665–1679, 2001.

[13] B. J. Blair, S.-H. Poon, and S. J. Taylor. Forecasting s&p 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. In *Handbook of Quantitative Finance and Risk Management*, pages 1333–1344. Springer, 2010.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[15] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[16] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[17] J. Boudoukh, R. Feldman, S. Kogan, and M. Richardson. Which news moves stock prices? a textual analysis. Technical report, National Bureau of Economic Research, 2013.

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

[19] D. Cabrera-Paniagua, C. Cubillos, R. Vicari, and E. Urra. Decision-making system for stock exchange market using artificial emotions. *Expert Systems with Applications*, 42(20):7070–7083, 2015.

[20] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, 2010.

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

[22] J. Y. Campbell, A. W. Lo, A. C. MacKinlay, and R. F. Whitelaw. The econometrics of financial markets. *Macroeconomic Dynamics*, 2(04):559–562, 1998.

[23] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211, 2016.

[24] E. O. Charles, G. J. Suci, and P. H. Tannenbaum. The measurement of meaning. *Urbana: University of Illinois Press*, 1957.

[25] H. Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[26] Z. Da, J. Engelberg, and P. Gao. The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32, 2015.

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

[28] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[29] A. K. Davis, J. M. Piger, and L. M. Sedor. Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Federal Reserve Bank of St. Louis Working Paper Series*, (2006-005), 2006.

[30] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, pages 703–738, 1990.

[31] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, pages 703–738, 1990.

[32] E. Demers and C. Vega. *Soft information in earnings announcements: News or noise?* Federal Reserve Board, 2008.

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

[34] W. Enders. *Applied econometric time series*. John Wiley & Sons, 2008.

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

[36] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[37] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[38] E. F. Fama. The behavior of stock-market prices. *Journal of business*, pages 34–105, 1965.

[39] S. Feuerriegel, M. W. Lampe, and D. Neumann. News processing during speculative bubbles: Evidence from the oil market. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 4103–4112. IEEE, 2014.

[40] S. Feuerriegel and D. Neumann. News or noise? how news drives commodity prices. In *Proceedings of the International Conference on Information Systems, ICIS 2013, Milano, Italy, December 15-18, 2013*, 2013.

[41] G. P. C. Fung, J. X. Yu, and W. Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402. IEEE, 2003.

[42] D. Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[43] T. Geva and J. Zahavi. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision support systems*, 57:212–223, 2014.

[44] L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801, 1993.

[45] G. Gorton and K. G. Rouwenhorst. Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2):47–68, 2006.

[46] Z. S. Harris. Distributional structure. *Word*, 1954.

[47] R. P. Hart. Redeveloping diction: theoretical considerations. *Progress in communication sciences*, pages 43–60, 2001.

[48] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[49] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[50] E. Henry. Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*, 3(1):1–19, 2006.

[51] M. Hu and B. 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.

[52] N. Jegadeesh and D. Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729, 2013.

[53] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan. Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1470–1473. ACM, 2013.

[54] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[55] S. Kelly and K. Ahmad. Sentiment proxies: computing market volatility. In *Intelligent Data Engineering and Automated Learning-IDEAL 2012*, pages 771–778. Springer, 2012.

[56] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.

[57] H. D. Lasswell. *Politics: Who gets what, when, how*. P. Smith New York, 1950.

[58] F. Lechthaler and L. Leinert. *Moody Oil-What is Driving the Crude Oil Price?* Eidgenössische Technische Hochschule Zürich, CER-ETH-Center of Economic Research at ETH Zurich, 2012.

[59] D. Leinweber and J. Sisk. Event driven trading and the'new news'. *Journal of Portfolio Management*, 38(1), 2011.

[60] G. Leng, G. Prasad, and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. *Neural Networks*, 17(10):1477–1493, 2004.

[61] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, pages 4–15. Springer, 1998.

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

[63] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on*, 38(6):1842–1845, 1992.

[64] B. Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[65] B. Liu and J. J. McConnell. The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? *Journal of Financial Economics*, 110(1):1–17, 2013.

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

[67] T. Loughran and B. McDonald. Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics*, 109(2):307–326, 2013.

[68] T. Loughran and B. McDonald. The use of word lists in textual analysis. *Available at SSRN 2467519*, 2014.

[69] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[70] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.

[71] G. Miller and C. Fellbaum. Wordnet: An electronic lexical database, 1998.

[72] G. Mitra and L. Mitra. *The handbook of news analytics in finance*, volume 596. John Wiley & Sons, 2011.

[73] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo. Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1):306–324, 2015.

[74] W. K. Newey and K. D. West. Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787, 1987.

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

[76] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[77] A. Ortony. *The cognitive structure of emotions*. Cambridge university press, 1990.

[78] C. D. Paice. An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–50. Springer-Verlag New York, Inc., 1994.

[79] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[80] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[81] B. Pang, L. Lee, and S. 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.

[82] D. Peramunetilleke and R. K. Wong. Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, 24(2):131–139, 2002.

[83] M. F. Porter. Snowball: A language for stemming algorithms, 2001.

[84] R. Quirk, D. Crystal, and P. Education. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press, 1985.

[85] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[86] J. C. Sager. *A practical course in terminology processing*. John Benjamins Publishing, 1990.

[87] J. C. Sager, D. Dungworth, P. F. McDonald, et al. *English special languages: principles and practice in science and technology*. John Benjamins Pub Co, 1980.

[88] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464, 2012.

[89] R. J. Shiller. Measuring bubble expectations and investor confidence. *The Journal of Psychology and Financial Markets*, 1(1):49–60, 2000.

[90] C. A. Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48, 1980.

[91] E. Soyer and R. M. Hogarth. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3):695–711, 2012.

[92] P. J. Stone, D. C. Dunphy, and M. S. Smith. The general inquirer: A computer approach to content analysis. 1966.

[93] C. Strapparava, A. Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[94] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

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

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

[97] P. C. Tetlock, M. SAAR-TSECHANSKY, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[98] I. Titov and R. T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer, 2008.

[99] R. S. Tsay. *Analysis of financial time series*, volume 543. John Wiley & Sons, 2005.

[100] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[101] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[102] P. D. Turney, P. Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[103] H. R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, pages 3–27, 2014.

[104] F. Wex, N. Widder, M. Liebmann, and D. Neumann. Early warning of impending oil crises using the predictive power of online news stories. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 1512–1521. IEEE, 2013.

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

[106] J. Wiebe. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740, 2000.

[107] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[108] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[109] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang. Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 3, pages 2720–2725. IEEE, 1998.

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