

**Moral Sentiment: Investigating the Roles of
Ethics and Affect in Determining Asset Returns**

Jason Andrew Cook

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

School of Computer Science & Statistics

Trinity College, University of Dublin

September 2017

Declaration

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

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

Jason Andrew Cook

Summary

Research in behavioural finance suggests that there is a significant relationship between affective language in text and changes in market prices. Often this affective language has ethical connotations: words like ‘good’ and ‘bad’, for instance, are used both ethically and evaluatively. This thesis aims to explore this relationship between ethical and affective language. Specifically, it examines their respective roles in determining asset prices, and investigates whether ethical judgements can be substituted for the negative terms typically used in studies on sentiment analysis. To this end I develop a sentiment analysis system to construct a proxy of investor sentiment from unstructured texts, and subsequently uses this sentiment proxy for econometric modelling. This system is able to handle a variety of unstructured textual and financial data at various temporal frequencies. At its core the system comprises two primary components:

1. A textual analysis component, which takes unstructured textual data as input, pre-processes and parses this input, and produces an ethical and affect time series as output.
2. A statistical modelling component, which takes the sentiment time series and market data as input, and computes a statistical model which examines the relationship between the ethical and affect time series and market data.

In this thesis I adopt a vector autoregressive framework, which posits that there is a linear relationship between the ethical and affect time series and between market returns. This model also includes past values for returns, along with past values for other market variables known to account for various other market frictions. In addition, I also conduct rolling regression models, which involve partitioning the data-set into consecutive sub-samples and computing a new regression model in each sub-sample. This allows the user to test the stability of the model’s parameters over time, and to

determine whether the strength of the relationship between ethical/evaluative terms and returns varies over time.

To test whether ethical terms have any impact on asset pricing I conduct three major case studies at different levels of market granularity: a firm-level study (using news and market data about the infamous Enron), an industry-level study (using news and market data from the pharmaceutical industry) and a market-level study (using general financial news and returns from the DJIA index). The main findings from these studies are as follows:

- Overall a relationship was found to exist between negative affect and returns, and between vice terms and returns. The nature and extent of the relationship is determined by the market data, textual corpus and time period of study. In general negative or unethical news on the previous day impacted returns on the following day, implying that news was promptly incorporated into changes in share price.
- The relationship between negative/vice terms and returns appears to vary over time. This suggests that the influence of news on financial returns is confined to certain periods.
- For all three case studies it was found that there was variability in financial returns that could be explained using vice terms, but not using negative terms. Likewise, there was variability in financial returns that could be explained using negative terms, but not vice terms. This suggests that both the negative and vice variables contain useful information about returns that is not contained in only one of the variables. Thus it is not possible to simply substitute one variable with the other.
- General off-the-shelf sentiment dictionaries sometimes need adapting for a particular domain. For instance, the General Inquirer dictionary in its native form is not well-equipped to deal with texts in the pharmaceutical domain. Words like ‘cancer’, ‘disease’ and ‘symptom’ which the dictionary treats as negative (and vice) are not typically evaluative (or ethical) in this domain. Attempting to use a sentiment dictionary in a different domain from the one in which it was constructed can sometimes lead to unreliable results, and so sentiment dictionaries may need to be adapted.

Acknowledgements

This thesis could not have been completed without the help of my mentors, colleagues, family and friends.

To my supervisor Professor Khurshid Ahmad I express my sincere thanks for his wisdom, help and guidance. Professor Ahmad kept me on track whenever I strayed too far. He also taught me to ‘look at what the numbers are telling me’ - even when they went against my own preconceptions.

To my parents, Mark and Debbie, I am utterly indebted. Thank you for teaching me the value of education, and for your constant love and support. It’s been a bumpy ride these last few years, but your frequent Skype calls, spontaneous visits, and much needed encouragement kept me going whenever I was feeling low.

I am grateful to all my friends and colleagues, however there are some who deserve a special mention. I am particularly grateful to Stephen Kelly, Zeyan Zhao, Xiubo Zhang, Marian Reeves, Shane Finan and Megan Vigor for your advice, friendship and countless coffee breaks throughout the duration of this Ph.D.

A special thanks also goes to my partner Katherine Morales. She has been my rock throughout the duration over my thesis, listening to me and supporting me, even while trying to balance her own Ph.D. work. I couldn’t have finished this thesis without her support and will forever be appreciative.

I would also like to take this opportunity to express my thanks to Dexter Dunphy and Edward Kelly for their advice on the General Inquirer dictionary and disambiguation rules.

Finally, I would like to thank Science Foundation Ireland and the ADAPT centre (www.adaptcentre.ie) at Trinity College Dublin for generously funding this research. If not for their grant (#07/CE/I1142) this research would not have been possible.

Contents

1	Introduction	14
1.1	Introduction	14
1.2	Motivation	17
1.3	Research Questions	18
1.4	Main Findings	19
1.5	Contributions	20
1.6	Publications	22
1.7	Structure	22
2	Background and Literature Review	24
2.1	Introduction	24
2.2	Background	26
2.2.1	The Efficient Market Hypothesis	26
2.2.2	Behavioural Finance: The Potential Case for Sentiment	28
2.2.3	Sentiment Analysis & Evaluative Language	31
2.2.4	Computing Sentiment: The General Inquirer	33
2.3	Construction of Sentiment Proxies	40
2.3.1	Dictionary-Based Approaches	40
2.3.2	Machine Learning Approaches	47
2.3.3	Advantages and Disadvantages	49
2.4	Related Work in Sentiment Analysis	50
2.5	Information Sources	51
2.5.1	News	55
2.5.2	Corporate Disclosures	57
2.5.3	Social Media	60

2.6	Modelling & Statistical Estimation	62
2.6.1	Linear Regression Models	64
2.6.2	Autoregression	66
2.6.3	Vector Autoregression	67
2.6.4	Rolling Regression	69
2.6.5	Panel Regression & Event Studies	72
2.7	Key Market Variables	73
2.8	Summary	77
3	Methods	79
3.1	Introduction	79
3.2	Textual Analysis	81
3.2.1	Textual Data	82
3.2.2	Pre-processing	86
3.2.2.1	Duplicate Detection	87
3.2.2.2	Sentence Splitting	88
3.2.2.3	Named Entity Recognition	90
3.2.2.4	Tokenization	91
3.2.3	Sentiment Extraction	92
3.2.3.1	Lookup & Stemming	93
3.2.3.2	Disambiguation	94
3.2.4	Document Aggregation	97
3.3	Model Estimation and Evaluation	98
3.3.1	Market Data	99
3.3.1.1	Returns	100
3.3.1.2	Measures of Market Liquidity and Risk	101
3.3.2	Transformations	103
3.3.3	Data Alignment	104
3.3.4	Model Estimation	106
3.3.4.1	VAR Model	106
3.3.4.2	Model Assumptions	109
3.3.4.2.1	Stationarity	109
3.3.4.2.2	Uncorrelated Residuals	111

3.3.5	Model Evaluation	112
3.3.5.1	Individual Variable Tests	112
3.3.5.2	Joint Significance Tests	113
3.3.5.3	Mean Square Error	115
3.4	Summary	116
4	Results	118
4.1	Introduction	118
4.2	Case Study 1: Sentiment at the Market-Level	120
4.2.1	Introduction	120
4.2.2	Data	121
4.2.3	Methods	122
4.2.4	Model Estimation	124
4.2.5	Experimental Results	127
4.2.5.1	The Impact of Negative Terms on DJIA Returns	127
4.2.5.2	The Impact of Vice Terms on DJIA Returns	131
4.2.5.3	Exploring the Time-Varying Impact of Sentiment . . .	133
4.2.6	Summary of the Market-Level Study	136
4.3	Case Study 2: Sentiment at the Firm-Level	137
4.3.1	Introduction	137
4.3.2	Data	139
4.3.3	Methods	142
4.3.4	Experimental Results	143
4.3.4.1	Descriptive Statistics	143
4.3.4.2	The Impact of Negative and Vice Terms on Enron's Returns	144
4.3.4.3	The Impact of Returns on Sentiment	148
4.3.4.4	Exploring the Time-Varying Impact of Sentiment . . .	150
4.3.5	Summary of the Enron case study	152
4.4	Case Study 3: Sentiment at the Industry-Level	153
4.4.1	Introduction	153
4.4.2	Data	154
4.4.3	Methods	158

4.4.4	Experimental Results	167
4.4.4.1	Impact of Negative and Vice Terms on ARCA Returns	167
4.4.4.2	Consistency Checks	169
4.4.4.3	The Impact of Returns on Sentiment	171
4.4.4.4	Exploring the Time-Varying Impact of Sentiment on Returns	172
4.4.5	Summary of Industry-Level Case Study	174
4.5	Summary	176
5	Conclusion	180
5.1	Main Findings	182
5.2	Limitations & Future Work	186
	Appendices	204
A	GI's Contextual Disambiguation Rules	205
B	Stemming Algorithm	208
C	Durbin-Watson Statistics	213
C.1	Enron Case Study	213
C.2	Pharmaceutical Case Study	214
D	GI Negative and Vice Terms	215

List of Figures

1.1	Overview of the sentiment analysis system.	16
2.1	Overview of rolling regression	70
3.1	Core architecture of the sentiment analysis system.	80
3.2	System diagram of textual analysis procedure.	81
3.3	System diagram of statistical modelling pipeline.	99
3.4	NYSE volume between 2004 and 2015.	104
4.1	DJIA returns between 1984 and 1999	122
4.2	Rolling regression of negative and vice terms against DJIA index returns.	136
4.3	Enron's returns between 1997 and 2001	141
4.4	Rolling regression of negative and vice terms against Enron's returns. .	151
4.5	ARCA returns between 2007 and 2015.	159
4.6	Rolling regression of negative and vice terms against ARCA pharmaceu- tical returns.	173

List of Tables

2.1	Emotional and psychological biases that affect human decision making. .	30
2.2	Frequencies of moral and affect terms in the GI dictionary.	34
2.3	Frequency of negative and positive words in four major sentiment dic- tionaries.	46
2.4	Summary of pertinent literature grouped by text type	53
2.5	Summary of pertinent literature grouped by frequency of market data. .	75
3.1	UML diagram for the Token class	86
4.1	Different senses for the word ‘capital’ in the GI dictionary.	125
4.2	Regression of negative terms against DJIA returns.	128
4.3	Descriptive statistics in the ‘Abreast of the Market’ Column	132
4.4	Regression of vice terms against DJIA returns.	133
4.5	Composition of the Enron corpus	140
4.6	Augmented Dickey-Fuller tests for stationarity in the Enron dataset. . .	141
4.7	Descriptive statistics for the Enron corpus	144
4.8	Regression of negative and vice terms against Enron’s winsorized returns	145
4.9	Regression of negative and vice terms against Enron’s non-winsorized returns	147
4.10	Regression of Enron’s winsorized returns against negative and vice terms	149
4.11	Components of the ARCA Pharmaceutical Index.	157
4.12	Composition of the pharmaceutical corpus	158
4.13	Descriptive statistics for the ARCA Pharmaceutical Index	159
4.14	Augmented Dickey-Fuller tests for stationarity in the pharmaceutical dataset.	160
4.15	Top 20 most frequent negative word senses in the pharmaceutical corpus	163

4.16	Top 20 most frequent vice senses in the pharmaceutical corpus.	164
4.17	Top 30 most frequent terms in the pharmaceutical override dictionary. .	166
4.18	Regression of negative and vice terms against ARCA pharmaceutical returns	168
4.19	Regression of negative and vice terms against ARCA pharmaceutical returns without the domain override dictionary	170
4.20	Regression of ARCA pharmaceutical returns against negative and vice terms	172
4.21	Summary of key VAR Results.	176
4.22	Summary of rolling VAR models	177
A.1	Basic structure of the GI's contextual disambiguation rules.	206
A.2	Example disambiguation rules for the word 'kind'	207
C.1	Results of Durbin-Watson test for autocorrelation in the residuals (Enron Case Study)	214
C.2	Results of Durbin-Watson test for autocorelation in the residuals (Phar- maceutical Case Study)	214
D.1	GI negative and vice terms with frequency ranks in each corpus	216

List of Abbreviations

AofM	Abreast of the Market (opinion column)
AR	Autoregression
ARCA	NYSE ARCA Pharmaceutical Index
CBOE	Chicago Board Options Exchange
CMLR	Classical Multiple Linear Regression
DJIA	Dow Jones Industrial Average
GI	General Inquirer
IPO	Initial Public Offering
L&M	Loughran & McDonald (e.g. Loughran & McDonald (2011))
MSE	Mean Square Error
NBER	National Bureau of Economic Research
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares
POS	Part-of-Speech
S&P	Standard & Poor
SEC	Securities & Exchange Commission
VAR	Vector autoregression
VIX	Volatility Index (CBOE)
WSJ	Wall Street Journal

Chapter 1

Introduction

1.1 Introduction

Financial markets, like any market, are ostensibly driven by a supply and demand equilibrium involving groups of buyers and sellers. As demand increases, and supply diminishes, the amount that buyers are willing to pay for a particular security (shares, bonds, currencies and so on) also increases. As demand decreases, an abundance of supply causes sellers to sell the security at a lower cost. At any given time the price of a security reflects both the amount that buyers are willing to pay for an item, and the amount that sellers are willing to sell it for. The goal of investment or trading is to attempt to buy an undervalued security, with the hope that the demand-supply equilibrium shifts and that profit can be made from a later resale of the asset.

Traditional financial theory would have one believe that markets are *efficient*, in the sense that the price of a security reflects all known information known about that security. Therefore a security cannot be underpriced or overpriced, as the price will instantaneously adjust to reflect any new information. However, this theory has been criticised by behavioural economists who argue that it does not take into account the different types of traders. While it may be true that institutional (professional) investors have access to similar information sources, there are also a wealth of individual investors in the marketplace which may not. For instance, there are a multitude of large news vendors (such as Dow Jones News Service, Thompson Reuters, Bloomberg) that charge large subscription-fees for access to instantaneous market information, which is not something to which an average individual investor would necessarily have access. This results in an *information asymmetry* where one group has access to more infor-

mation than another, and can therefore exploit market inefficiencies. However, even if access to news sources was identical for all parties there is no guarantee that investors or traders will digest the news in the same way. Some investors are presumably more efficient at processing the information they read, and some have more experience in determining the information's impact.

It has also been observed that investors and traders can, at times, be influenced by emotions and other psychological biases (Shiller (2000)). For example, Kahneman & Tversky's Prospect Theory (Kahneman & Tversky (1979)) would have one believe that people feel losses more than gains. This can lead to investors holding onto losing stocks in the hope that the asset price reverses. Investor overconfidence has also been the detriment of a number of investors (Barber & Odean (2011)), who believe that they are better than others in their ability to time their buying and selling decisions. Additionally, herd behaviour - the tendency for individuals to 'follow the crowd' - has also been observed in financial markets, most noticeably evident in stock market bubbles and crashes. It has been suggested that so-called investor sentiment plays a key role in the formation of market bubbles (e.g. Shiller (2000)), which invariably lead to market collapse.

Investors, then, can be driven by their emotions and other psychological biases. These emotions can cause them to overvalue or undervalue a particular instrument relative to its 'true' value, ignoring the fundamental forces that shape the value of the asset. By analysing the sentiment in the market, it may be possible to glean an insight into investors' thoughts and feelings at the aggregate level. This information could then be used to determine whether an asset's current market value is overpriced or underpriced.

In this thesis I therefore develop a system to collect, aggregate and extract information from unstructured text (see Figure 1.1). This system (to be discussed in Section 3.1) constructs a proxy of investor sentiment by analysing the evaluative language contained in text, and subsequently uses this proxy in an econometric model to explore the effect of investor sentiment on financial returns. While the system is not designed with automated trading in mind, it could be used as a decision support system to help mitigate risk. However, the system's main goal is as a research tool to explore alternative measures of investor sentiment. Traditionally, sentiment analysis systems measure the degree of positive or negative tone in a document. However, much of the positive

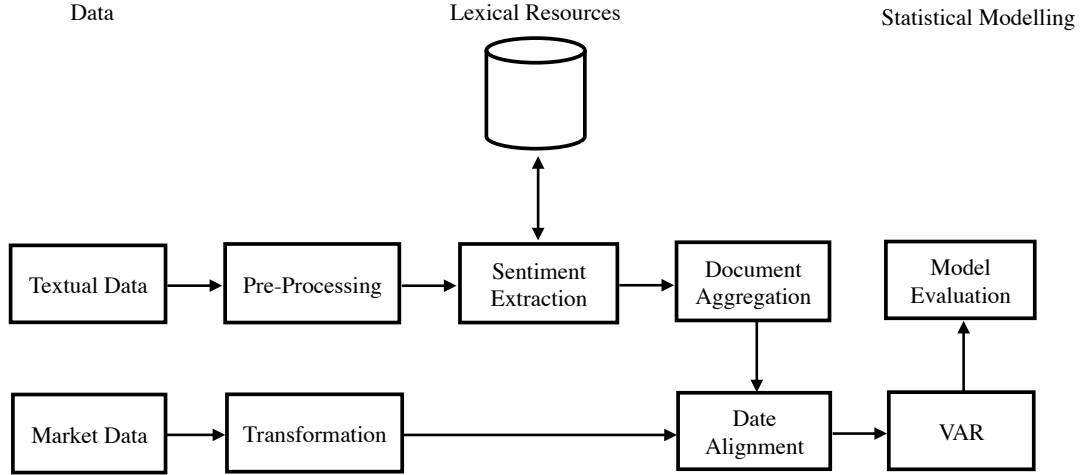


Figure 1.1: Overview of the sentiment analysis system. The term ‘VAR’ is an abbreviation of ‘Vector Autoregression’ - one of the main statistical models used in this thesis.

or negative judgements in texts tend to carry *ethical* connotations. The goal of this thesis, therefore, is to explore the effect of ethical judgements on financial markets.

I conduct three case studies to evaluate the system, and analyse the relationships between the ethical and affect terms. In the first I use the Dow Jones Industrial Average index to explore the effect of ethical and affect terms at the market-wide level. In the second I use the erstwhile Enron corporation to explore the effect of ethical terms on firm-level returns in the domain of financial and accounting scandals. And in the third I conduct a case study using textual and industry-level market data from the pharmaceutical industry. Collectively these case studies show the ability of the system to handle a wide range of market and textual data, at different levels of granularity and during different time periods.

Overall in this study I find a strong, yet complex relationship between ethical language and investor sentiment. An increase in negative or unethical ‘vice’ terms on the previous day generally leads to downward pressure on financial returns the following day. However, the strength and extent of the relationship is found to vary over time, and also depends upon the textual data being used and the industry being investigated. Both negative and vice terms are also observed to contain useful information about financial returns that is not found in using only one of the variables in isolation.

1.2 Motivation

Research in sentiment analysis has historically constructed sentiment measures by analysing the positive or negative evaluative language in texts. However, by inspecting dictionaries of sentiment-bearing words and phrases (hereafter ‘sentiment dictionaries’) it is evident that many of these entries also carry ethical connotations. Indeed the motivation for this study first arose following a careful examination of the General Inquirer dictionary (Stone et al. (1966), Stone (1969), Kelly & Stone (1975)), which has been widely used in the sentiment analysis literature (e.g. Tetlock (2007), Tetlock et al. (2008), Engelberg (2008), Cook & Ahmad (2015), Feldman et al. (2010), Kothari et al. (2009), Twedt & Rees (2012), Heston & Sinha (2016), Das & Chen (2007), Huang et al. (2014), Ferguson et al. (2015), Chen et al. (2014)). A large overlap is observed between the dictionary’s affect categories of positive and negative and the dictionary’s ethical categories of virtue and vice, with the ethical words and phrases comprising a subset of the affect terms. This led me to wonder whether the sentiment proxies constructed to measure affective language in text are really measuring ethical language. In this thesis I am therefore concerned with whether this ethical subset of terms can be used as a proxy for investor sentiment in place of the traditionally used categories of negative affect.

The work in this thesis is situated during a time when there is an increasing focus on ethical business practices and corporate social responsibility. Following a wave of corporate and accounting scandals towards the end of the previous millennium, a number of legislative changes have been made to ensure that businesses are acting in a responsible manner¹. Businesses are no longer considered to have only (financial) responsibilities to their stakeholders, but are also expected to have ethical responsibilities towards their employees, their organization, their customers, local communities, the environment and society at large. Issues of ethics are incorporated into all aspects of business decision-making, from avoiding discrimination in the workplace (providing equal opportunities regardless of ethnicity, sex, age, etc.) to ensuring appropriate handling of personal data. In addition, most large businesses today have various codes of ethics prescribing the difference between ‘right’ and ‘wrong’ behaviour. These codes of

¹An example of one such legislative change is the United States’ Sarbanes-Oxley Act (2002), introduced to ensure organizations produce more accurate financial statements

ethics can be implemented at an organizational level (describing how the organization makes ethical decisions), an employee level (describing how employees should or should not behave) or an industry level (defining how all organizations within a given industry should behave).

Today we live in an age where every business decision faces intense scrutiny from the media. With the rise of social media and other user-generated content, anyone can instantaneously publish new information on the Web. Traditional news vendors often feature stories about recent scandals, which can cause investors to react (or overreact) leading to changes in a company's share price. Sometimes these changes are temporary, as was the case with the September 2015 Volkswagen emissions scandal, where the company's share price initially decreased but was followed by a subsequent reversal. However, sometimes these changes can be permanent, as was the case with the 2001 Enron scandal, where the company's share price declined and the company was subsequently forced into bankruptcy.

Given, then, that ethical reports of business misconduct in news can have temporary or long-lasting effects on the stock value of a company, one wonders whether a systematic approach can be taken to analyse this news automatically and make decisions that could be used by investors and traders. In this thesis I develop such a system, interleaving work from various disciplines including behavioural psychology, linguistics, computer science, econometrics and finance.

1.3 Research Questions

In this thesis I investigate the relationship between ethics and affect, and consider their respective roles in determining asset prices. To do so, I intend to answer the following questions:

- Are affect terms rooted in ethical judgements?
- Do ethical vice terms impact market prices in the same manner as negative affect terms? Can these ethical terms be substituted for negative terms to derive a proxy for investor sentiment?
- Does the impact of negative affect/vice terms vary over time?
- Is the impact of negative/vice terms confounded by domain-specific vocabulary?

To answer these questions I present three case studies at different levels of market granularity: the individual firm-level, the industry-level and the market-level. Each of these studies spans a different time period and utilises texts from different sources, allowing me to explore the impact of ethical evaluation under a variety of different conditions.

1.4 Main Findings

This research reveals that there is a strong yet complex relationship between ethical and affective language. Three major case studies are presented in which both ethical and affective terms are used to derive a sentiment proxy. A series of econometric models are subsequently computed which examine the relationship between the derived proxy and financial returns. Overall, it is observed that both negative and ethical terms impact financial returns, but the extent of this impact depends upon the market data and period of study. Generally the impact of news on financial returns is strongest on the previous day: a unit increase in the number of negative terms on the previous day leads to a -4.4 to -10.1 basis point² decrease in next day returns: conversely a unit increase in the number of vice terms leads to a -1.6 to -13.4 basis point decrease in next day returns. This next-day effect implies that share prices quickly change to reflect changes in investor sentiment.

It is also observed that the relationship between the sentiment proxy and financial returns varies over time. For instance, the nature of the relationship is observed to be different during recessionary periods than during expansionary periods. A series of rolling regression models are computed for each of the three case studies, and reveals that the strength of the relationship between affect terms, ethical terms and financial returns is confined to particular time periods. During these periods the negative and vice terms explain variability in returns that cannot be explained by other market variables.

This research also found that general sentiment dictionaries sometimes need to be adapted for a particular domain. Typically, sentiment dictionaries contain a list of terms that are annotated (either manually or computationally) with their a priori

²A basis point is equal to 1/100th of a percent, and is a common measure of percentages in finance.

polarity. However, attempting to use a sentiment dictionary in a different domain from the one in which it was constructed can sometimes lead to unreliable results. Specifically, the evaluative terms defined in the sentiment dictionary may not have evaluative connotations in a particular domain. This research found that the General Inquirer sentiment dictionary needs to be adapted for use in the pharmaceutical domain, as words such as ‘cancer’, ‘disease’ and ‘symptom’ (which the General Inquirer tags as ‘negative’) are generally not used in an evaluative manner in the pharmaceutical domain.

1.5 Contributions

The main contribution of this thesis is the research into the effect of ethical language on financial markets. Typically, research on sentiment analysis tends to measure the number of affective words in text, which are then used as a proxy for investor sentiment in econometric models. However, in this thesis I investigate a subset of these affect terms - the ones rooted in ethical values - to discover whether they contain useful information about financial returns. I compare their impact to the impact of negative affect terms, to see whether affect terms are rooted in ethical judgements. In particular, I examine whether this subset of ethical terms can be used in place of the traditionally used affect terms. To this end I construct a text analysis system that takes unstructured texts as input, and converts these texts into ethical and affect time series that can be used in econometric modelling.

The second major contribution of this thesis is the design and implementation of the system itself. The system designed in this thesis is capable of working with any unstructured textual data, at any temporal frequency. Using the General Inquirer dictionary and its associated disambiguation procedures (Stone et al. (1966), Kelly & Stone (1975)), the system extracts ethical and evaluative information from the texts, and creates a quantitative sentiment time series as output. This sentiment time series is then aggregated with financial market data, and statistical models are subsequently computed.

One innovative component of the system is the override component, which is designed to overcome issues of domain-specific vocabulary when using a general sentiment dictionary. In certain domains evaluative terms in a general sentiment dictionary may

not be evaluative in the target domain. For instance, the word ‘cancer’ appears as negative (and therefore vice) in the General Inquirer dictionary, but would not typically be used in an evaluative manner in the pharmaceutical domain. To overcome this issue I introduce the notion of a domain-override dictionary, containing a list of lexical items from a particular domain. The system is instructed to ignore these items, and not to apply any evaluative tags from the GI dictionary.

To evaluate this system I analyse the effect of ethical and affect terms in three major case studies:

1. In the first study I analyse the ability of ethical and affect terms to explain variability at the index-level. To this end I use the Dow Jones Industrial Average as a proxy for market value.
2. In the second case study I extend this research to explain variability at the firm-level. Specifically, I investigate the role of ethical and affect terms in explaining the returns of Enron.
3. In the third study I explore the role of sentiment in explaining returns at the industry-level. Using news about 25 companies in the pharmaceutical domain, I explore the impact of ethical and affect terms on a major pharmaceutical index.

Each case study utilises a distinct textual corpus comprising different sources. Each study also involves a distinct time period, encompassing different business cycles and a different macroeconomic structure. Collectively, these three case studies demonstrate the ability of the system to handle a wide range of market and textual data.

In this thesis I also provide a detailed overview of the composition of the General Inquirer dictionary and its associated disambiguation rules. The General Inquirer dictionary has been used extensively for performing sentiment analysis in financial markets. However, typically researchers are unaware of its procedures for handling polysemy. As a result, Chapter 2 provides a synthesis of its features, which would be of benefit to other researchers who wish to use the dictionary in their own studies.

1.6 Publications

The work in this thesis builds upon the following publications:

Cook, J. A., & Ahmad, K. 2015. Behaviour and Markets: The Interaction Between Sentiment Analysis and Ethical Values? In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 551-558). Springer International Publishing.

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

In Cook & Ahmad (2015) we presented initial evidence of the relationship between ethics and affect. Using the former Enron as a case study, we explored the association between ethical terms, affect terms, and financial returns in a large corpus of reportage and other news-related texts. In Chapter 4 I expand upon this case study further by employing a vector autoregressive framework to examine the impact of negative and vice terms on financial returns.

In Cook et al. (2016) we investigated the statistical properties of negative sentiment across a variety of different corpora (both general language and special language). The main finding from this paper was that while the frequency distribution of negative terms tends to vary between corpora, the frequency distribution within corpora is relatively stable even across different corpus sizes. Thus extending the number of documents in each corpus had little effect on the underlying distribution of the negative terms.

1.7 Structure

The remainder of this thesis is structured as follows.

In Chapter 2 I present an overview of background work and review the related literature on sentiment in finance. I begin with a discussion of efficient market theory, before motivating the role of investor sentiment in the market place. I then define

sentiment and discuss the various ways in which it can be computed, together with advantages and disadvantages of each approach. I then discuss some of the main econometric models used, before ending the chapter with a summary of the key findings to date.

In Chapter 3 I present an overview of the system’s architecture, discussing the process of procuring and cleaning data, transformations, and extracting a quantitative sentiment proxy from unstructured text. I then discuss various statistical processes that are used in this thesis to model the effects of sentiment.

In Chapter 4 I conduct 3 case studies using the system outlined in Chapter 3. These case studies are chosen to demonstrate the variety of market and textual data which the system is capable of dealing with. In the first case study I explore the effect of evaluative and ethical news on financial returns at the market-level, using the Dow Jones Industrial Average (DJIA) index as a proxy for market value. In the second case study I use the infamous Enron corporation, which was involved in a series of accountancy and financial scandals during the late 1990s and early 2000s. In the third case study I use news about 25 companies from the pharmaceutical industry over the period 2007 through 2015, and examine the effect of evaluative and ethical terms in the news on returns of the NYSE’s ARCA Pharmaceutical index.

In Chapter 5 I conclude and summarise the major findings of this thesis, and identify some of the limitations of this study. I end with a series of possible suggestions for future work.

Chapter 2

Background and Literature Review

2.1 Introduction

Financial markets are full of market participants, each driven by their own objectives, attitudes and emotions. The common goal of these participants is simple: to maximize wealth while simultaneously minimizing risk. However, despite this common goal, financial markets often behave as if they were driven by a random process. It seems as if the market is full of ‘animal spirits’ (Keynes (1936)).

In order to gain a competitive edge, market participants often turn to technical tools to analyse a market’s dynamics. These tools are designed to guide investors or traders as to the general direction of the market, signalling when it is safe to enter or exit a market, and indicating when the market’s conditions are about to change. One such set of tools are *sentiment indicators*, designed to monitor the mass opinions of market participants. The purpose of these ‘sentiment indicators’ is to analyse when market participants are feeling confident or optimistic (i.e. ‘bullish’) and when they are feeling pessimistic or afraid (i.e. ‘bearish’). Of course it is not possible to analyse the opinion of every market participant at once, and even if it were this would be likely to change in an instant. As a result it is not possible to directly measure investor sentiment, and so market participants turn to indirect measures in the form of sentiment proxies.

One set of sentiment proxies is sentiment surveys, which aim to sample individual market participants and see how they feel the stock market will perform over the forthcoming months. For instance, the American Association of Individual Investors (AAII)

survey (www.aaii.com/sentimentsurvey) is a weekly survey that polls its members to find out whether they believe the market will be bullish, bearish or neutral over the next six months. Another common survey is the Michigan Consumer Sentiment Index (MCSI) - a monthly sentiment index published by the University of Michigan. The MCSI is constructed following monthly telephone interviews, with each person being asked fifty core questions about the overall economy. A third sentiment proxy is the Westpac-Melbourne Institute Survey of Consumer Sentiment. This sentiment proxy is essentially an average of five indices which survey consumers on their financial situation over the past and forthcoming year, their anticipated economic conditions over the next five years, and whether now is a good time to buy major household items.

However, conducting sentiment surveys in the manner previously described is both timely and costly. In an age where computers can submit trading orders in a fraction of a second, any weekly or monthly sentiment survey is likely to be out of date by the time it is published. Consequently, what is needed is a computational approach to the problem, which avoids the manual burden of conducting surveys and aggregating results. In this thesis I propose such a system.

The goal of a sentiment analysis system is to take unstructured textual data as input (such as news, magazines, corporate documents, social media), and compute the overall sentiment in the data. It is believed that textual sentiment functions as a proxy for investor sentiment in the market place, with negative and positive evaluative language proxying for bearish and bullish behaviour respectively. If each document is tagged with temporal information then a sentiment time series can subsequently be produced. This sentiment time series can then be used as a variable in econometric models to explore the relationship between sentiment and key economic variables. In a practical sense, such a system could be used to mitigate risk or uncertainty or to predict key market movements. The question of course is how to compute this sentiment time series.

In this chapter I review the literature on sentiment analysis in financial markets. Beginning with a discussion of traditional financial theory, I motivate the role of behavioural and psychological factors in the market place. I then go on to define sentiment, and discuss the main approaches to computing it. Finally, I end with a discussion of related work on sentiment analysis in the finance literature, discussing key relationships between sentiment and market variables. This latter discussion is divided by

both textual source and market data, being two of the inputs to the system that will be described in Chapter 3. I also discuss some of the main statistical approaches that have been used to date, and provide background information on the statistical methods that will be used in the remainder of this thesis. The main goal of this chapter is to situate my own study within the body of existing literature. By its end, the reader should have a good overview of existing approaches to sentiment analysis in financial markets, together with knowledge of typical data and methods that have been studied and some of the core findings to date.

2.2 Background

2.2.1 The Efficient Market Hypothesis

Over the last 50 years the efficient market hypothesis (EMH) has been a central tenet of financial theory. The EMH, popularised by the Nobel Laureate Eugene Fama (Fama (1970)) in the late 1960s, posits that financial markets are *informationally efficient*. At its core the theory implies that a market is efficient if its prices fully reflect all available information about an asset. Any new information would have already been quickly incorporated into the asset's price, and thus the asset price reflects the 'true' (i.e. fundamental) value. With no opportunities for buyers to buy 'underpriced' assets or for sellers to sell 'overpriced' assets, it is claimed that it is impossible for an average investor to consistently 'beat' the market except by chance. The motivation for testing the EMH comes from the work of economists who are generally in favour of so-called free markets - that is markets that are free from regulations and constraints. This is a complex and controversial topic and not quite the focus of this thesis. I am more interested in behavioural finance where the argument is that markets lack inherent efficiency in that hypotheses such as the EMH fail to take into account 'human behaviour'. This is discussed further in Section 2.2.2.

Fama (ibid.) asserts that there are three forms of efficiency to the model, which depend on the type of information incorporated into the asset's price. The weak-form efficiency model claims that the asset price reflects all information about the historical prices. Therefore, it is not possible to predict future prices by analysing historic prices, as future movements will follow a random walk. In the semi-strong form of the EMH, it is claimed that prices contain all publicly available information (e.g. companies'

announcements, annual earnings figures) as well as all historical share price information. New information is said to be instantly incorporated into asset prices in an unbiased fashion. Finally, in the strong-form efficiency model, asset prices are said to reflect all information, including insider information that is unavailable to investors and other members of the public. This suggests that in the long-run investors cannot gain a competitive edge from knowing confidential corporate information.

The concept of market efficiency can be expressed mathematically in Equation 2.1. Here, the expected return (change in asset price), given the set of all available information³, is equal to ε_t - a normally distributed random variable with a mean of 0. Thus the long-run expected return, given the set of available information, will be equal to 0.

$$\mathbf{E}(r_t | \{information\}) = \varepsilon_t \quad (2.1)$$

Where:

r_t = return (change in price) at time t .

ε_t = a random error term at time t (with an expected value of 0).

The EMH therefore discounts the role of investor sentiment, claiming that no market participants can gain a competitive edge by obtaining new information. It also assumes that investors all have access to the same information, that they are equally skilled in the processing of this information, and that prices fully reflect this information at all times. However, the EMH has received a number of criticisms in recent years, and there are a number of anomalies which the theory cannot explain. In the next section I discuss some of these anomalies, and motivate the role of investor sentiment in the market place.

³The set of information could either be historical price information, in the case of the weak-form EMH, public information in the case of the semi-strong form of the EMH or all public and private information in the case of the strong form of the EMH.

2.2.2 Behavioural Finance: The Potential Case for Sentiment

Efficient market theory assumes that market participants are rational decision makers, who have access to identical information and process this information in an identical fashion⁴. However, many researchers advocate for the presence of irrational ('noise') traders, who cause the market to be inefficient (e.g. Black (1986), De Long et al. (1990)). Black (1986) argues that if market participants all have access to identical information, then no trading would take place as all stocks would reflect their true, fundamental value. However, in reality investors look for opportunities to buy assets they believe to be undervalued, and sell assets they believe to be overvalued. The key word here is 'believe', as market prices are determined by perceptions about what the true value should be. Investors' differing beliefs are formed from different sources of information (Black (1986)). In some cases this information is not really information at all, but investors trading upon 'noise' as if it were information (ibid.).

The role of irrationality in human decision making has been emphasized in the literature on behavioural psychology. A number of scholars have suggested that investors are influenced by cognitive and psychological biases, such as loss aversion, framing, overconfidence, and recency bias. These biases, described in Table 2.1, are said to influence investor decision making, causing investors to overvalue or undervalue an asset beyond its fundamental value. To quote Thaler, 'it is necessary to entertain the possibility that some of the agents in the economy behave less than fully rationally some of the time' (Thaler (2005)).

Behavioural finance attempts to challenge the traditional efficient market paradigm, emphasizing the role of human behaviour in determining asset prices. Besides these behavioural biases, a number of market inefficiencies have been noted in the literature that dispute the EMH. For instance Shiller (2000) argued that the existence of speculative financial bubbles are unexplained by traditional efficient market backed theories. He cites the Great Crash of 1929, Black Monday in October 1987, and the dot.com bubble of the late 1990s/early 2000s as examples. The more recent US subprime mortgage crisis of 2007-2008 can also be added to this list. The key point noted by Shiller is that

⁴The strong form of the EMH suggests that investors cannot profit even from private, insider information, however this version of the EMH is not widely believed in practice, and a number of insider trading cases clearly contradict this version.

investors' 'irrational exuberance' causes them to overvalue a particular market, leading to speculative financial bubbles and ultimately to market collapse. During these bubble periods Equation 2.1 appears not to hold.

Some have argued that the EMH is weak on the notion of price momentum, where rising (falling) prices continue to rise (fall) over a prolonged period of time. This can occur, for instance, as a result of investors overreacting or underreacting to information (Daniel et al. (1998), DeBondt & Thaler (1985)), causing prices to be pushed further away from fundamentals. According to the EMH, there should be no overreaction or underreaction, as prices already contain all available information. It has also been claimed that past 'losers' (stocks with low, negative returns) tend to subsequently outperform past 'winners' (stocks with high, positive returns) (Jegadeesh & Titman (1993)). This claim implies a pattern to prices that can be predicted, violating even the strongest form of the EMH. As a result, behavioural finance approaches assume that there is systematic and predictive information contained in the error term in Equation 2.1 - information that can be used to explain variability in market returns.

Behavioural finance approaches therefore argue that sentiment is essential to investor decision making. At certain times it is argued that investors and traders succumb to 'irrational exuberance' that can cause them to overprice or underprice assets above or below their 'true', fundamental value. Evidence in behavioural finance suggests 'systematic and significant deviations from efficiency (that) are expected to persist for long periods of time' (Shleifer (2000)). The key point to note here is that while the EMH suggests that investors' emotions do not impact prices, behavioural finance argues that they do. Emotions are articulated in evaluative language through value judgements, as well as through the use of non-verbal signals. However, it is not clear in which direction the causality between emotion and prices lies. While it could be the case that price changes are reacting to value judgements, it is also plausible that value judgements are made in response to changes in prices. A feedback loop could also be possible whereby emotions influence prices which in turn influence emotions.

Table 2.1: Emotional and psychological biases that affect human decision making (see Kahneman & Tversky (1972), Kahneman & Tversky (1979), Shiller (2000)). The first column denotes the name of the bias. The second column denotes a description of the bias. The third column denotes its implications in finance, specifically its implication on efficient market theory. The fourth column denotes relevant references.

Bias	Description	Implications for Investor Behaviour	Ref.
<i>Loss Aversion</i>	People feel losses more than gains	Investors hold onto falling stocks, and sell rising stocks too quickly	Kahneman & Tversky (1979)
<i>Framing Effects</i>	Problems framed in different ways leads to different decisions	Information from different sources could lead to different interpretations	Tversky & Kahneman (1985)
<i>Overconfidence</i>	Overconfident beliefs in abilities or knowledge	Could lead to excessive risk-taking	Barber & Odean (2001)
<i>Representative bias</i>	Assuming future performance will resemble the past	Investors may buy rising stocks, expecting them to keep increasing	Kahneman & Tversky (1972)
<i>Confirmation bias</i>	Tendency to find information to validate one's prior beliefs or opinions.	Individual investors could interpret news differently depending on their prior beliefs	Nickerson (1998)
<i>Recency bias</i>	People tend to overweight recent information	Investors may more prominently recall or emphasize more recent events	Pompian (2012)
<i>Anchoring</i>	Become fixated on a particular information and using it to make decisions	Investors may anchor to a particular price point of what they think an asset's price <i>should</i> be	Tversky & Kahneman (1975)

2.2.3 Sentiment Analysis & Evaluative Language

The term ‘sentiment’ has its etymology in the Latin verb ‘sentire’, meaning ‘to feel’. Sentiment analysis usually refers to the acquisition and analysis of peoples’ feelings towards a particular entity or topic. In the context of finance and accounting, this would usually refer to investors’ feelings, and the target of their sentiment would be a particular security or financial market (Kim & Kim (2014)). Nowadays, the definition of sentiment has been extended to include the notion of opinions. In addition to its original Latinate definition, the Oxford English Dictionary defines sentiment as ‘an opinion or view as to what is right or agreeable’. This extended definition of sentiment thus includes both opinions and feelings towards a particular subject matter. Indeed the terms ‘sentiment analysis’ and ‘opinion mining’ are often used interchangeably (e.g. Liu (2012), Liu & Zhang (2012)), and the literature treats sentiment analysis and opinion mining as synonymous terms.

Opinions are arguably expressed through the use of evaluative utterances (Hunston (2010)). Although there are different linguistic theories on the meaning of evaluative language (e.g. *stance* (Biber & Finegan (1989), Jaffe (2009)) or *appraisal theory* (Martin & White (2003))), the focus of sentiment analysis research is typically on identifying positive and negative opinions following Osgood et al.’s (1957) work on the measurement of meaning.

Osgood and his colleagues (researchers in social psychology) proposed that the connotative meaning of words could be represented by a vector in multidimensional space; the dimensions of this space, they asserted, could be determined empirically. To this end they developed a psychological method called the semantic differential scale, designed to measure reactions to stimulus concepts. The semantic differential was essentially a 7-point Likert scale with a pair of bipolar adjectives at each end. Participants were presented with a concept and a set of semantic differential scales, and asked to rank where on the scales they would place the stimulus concept. In the example which follows the concept is ‘LADY’, and participants would be asked to place where on the scale of good-bad they would place this concept.

LADY

good __ : __ : __ : __ : __ : __ : __ *bad*

Each differential scale represents a particular dimension in a multidimensional semantic space. The position on the scale corresponds to the location of that concept in the space. Osgood and colleagues administered the test to subjects using a variety of concepts and scales, and subsequently used a factor analysis procedure to reduce the number of semantic dimensions. From this they concluded that the bulk of the meaning of words could be summarised by three dimensions: an evaluative dimension (i.e. how good or bad something is), a potency dimension (i.e. how strong or weak something is) and an activity dimension (i.e. how active or passive something is).

Of Osgood et al.’s (ibid) three affect dimensions, the evaluative dimension⁵ was said to contribute the most towards semantic variation. This latter point provides the basis for the computational modelling of sentiment, as it suggests that the study of polarized language use is possible by quantifying words and phrases along this evaluative dimension. Thus, lexical items that encode a desirable state (i.e. positive terms such as ‘good’ or ‘excellent’) would be located at one end of the evaluative scale, and items that encode an undesirable state (i.e. negative terms such as ‘bad’ or ‘terrible’) would be located at the other end, with the distance along the scale indicating the intensity of the evaluation. Sentiment analysis therefore amounts to deriving methods that can take qualitative linguistic data as input, and map it onto this evaluative dimension. In the majority of the sentiment analysis literature this is typically achieved by categorizing words or phrases as if they were positive, negative or neutral. Often this is considered a binary classification, although Osgood and colleagues made no such requirement in their original theory.

During the development of their theory, the authors in Osgood et al. (1957) noticed that evaluative words could be subdivided into four distinctive clusters: a ‘morally evaluative’ cluster (comprising words such as ‘sacred’, ‘fair’, ‘valuable’), an ‘aesthetically evaluative’ cluster (comprising words like ‘pleasant’, ‘sweet’, ‘sour’), a ‘socially evaluative’ cluster (comprising words like ‘beautiful’, ‘brave’, ‘rich’) and an ‘emotionally evaluative’ cluster (comprising words like ‘peaceful’, ‘relaxed’, ‘calm’). While this observation was ultimately not integrated into their theoretical framework it raises an interesting topic: perhaps it is the morally evaluative terms are the ones driving financial markets?

⁵The evaluative dimension includes bi-polar adjective pairs such as positive-negative, good-bad, beautiful-ugly.

There is a voluminous literature that explores the relationship between evaluative language and financial markets. The majority of this work focuses on the distinction between ‘positive’ and ‘negative’ words first proposed by Osgood et al. (ibid). However, based on this latter observation by Osgood’s colleagues it could be the case that it is a subset of these evaluative terms (the ethically evaluative terms) that are driving the results in previous studies. Consequently, in this thesis I explore whether there is a relationship between ‘morally evaluative’ terms and financial markets. In the next section I discuss how these morally evaluative terms can be computed.

2.2.4 Computing Sentiment: The General Inquirer

One of the earliest attempts at mapping language onto Osgood et al.’s affect dimensions came in the form of the General Inquirer (GI) system (Stone et al. (1966), Stone (1969), Kelly & Stone (1975)), which was initially developed in the 1960s by a group of social psychologists at Harvard University. The GI system’s main goal was to provide a computational approach for performing *content analysis*, defined by the authors as ‘any research technique for making inferences by systematically and objectivity identifying specified characteristics within text’ (Stone et al. (1966)). While manual content analysis techniques had a longer history, such as the analysis of propaganda techniques during the First World War (Lasswell (1927)), the GI system sought to remove the researcher subjectivity and lack of replicability associated with manual analysis. To do so the authors proposed that the characteristics of texts could be determined computationally via a lookup procedure in a pre-defined dictionary. In this dictionary words were mapped onto one or more content categories, whereby semantically equivalent words would be represented by the same concept. For each word in the input text, the system performs a dictionary lookup of that word, and tags each word with any applicable categories. The General Inquirer dictionary contains semantic categories from a wide variety of different fields, including economy, political, legal, military, hostile, pleasure and pain. Importantly for this study, are Osgood’s affect dimensions of evaluation, activity and potency, denoted by the bipolar GI categories of positive and negative, active and passive and strong and weak respectively.

The GI dictionary was constructed over a period of many years and underwent several revisions whereby new categories were added or removed, and words belonging to existing categories were subsequently revised. Its current form consists of 11,788

word senses and 183 semantic categories derived from research in social psychology. A word may belong to one or more of these semantic categories, and words within the same category are considered to contribute to the overall theme of that category. The assignment of words to category is binary in nature: a word is either mapped to a particular category or it is not⁶.

In addition to Osgood et al.’s affect categories, this study utilizes two other GI categories, namely the categories of vice and virtue. These vice and virtue categories are defined by Dunphy et al. (1974) as ‘Characteristics of persons, processes or objects generally regarded by society (a white middle-class American society) as vices (virtues) or misfortunes (good fortunes)’. Thus the vice and virtue categories represent a semantic dimension of morality, with the vice category denoting words of an unethical (‘immoral’) nature and the virtue category denoting words of an ethical (‘moral’) nature. The total number of vice and virtue terms and Osgood’s affect categories are shown in Table 2.2. A complete list of the vice and virtue terms are shown in Appendix D.

Table 2.2: Frequencies of moral and affect terms in the GI dictionary.

Dimension		Number of Words in GI
Evaluative	Negative	2,291
	Positive	1,915
Activity	Active	2,045
	Passive	911
Potency	Strong	1,902
	Weak	755
Morality	Virtue	719
	Vice	686

An interesting observation about the vice and virtue categories is that nearly all of the terms in these categories are also tagged as negative or positive respectively. In fact, 92.55% of the GI’s vice terms are also tagged as negative, and 84.14% are tagged as

⁶In an early version of the GI dictionary the categories of ‘positive’ and ‘negative’ were graded on a 3-point scale, with positive-1 indicating a stronger assignment than positive-2, and positive-2 indicating a strong assignment than positive-3. However, this gradation was removed in subsequent revisions as it was deemed too difficult for annotators to agree upon.

positive. This shows that the vice and and virtue terms in the GI dictionary constitute a subset of the negative and positive terms respectively. What is particularly interesting about this observation is that it mimics the findings of Osgood et al. (1957), who found that evaluative terms typically clustered into four groups (morally evaluative, aesthetically evaluative, socially evaluative and emotionally evaluative). Even though the work in Osgood et al. (1957) and Stone et al. (1966) was conducted independently the authors essentially arrived at the same conclusion: that evaluative language appears to have an ethical foundation. This brings us to the main premise of this thesis, which seeks to establish whether these morally evaluative terms can be used in place of the traditionally used negative affect terms in order to derive a proxy of investor sentiment. In the remainder of this section I further develop the discussion of the General Inquirer dictionary, which provides background information necessary for understanding the system architecture in Chapter 3.

The authors of the GI system recognized that a large number of inflectional entries would tend to share the same content categories (Kelly & Stone (1975)). Consequently, to reduce the number of duplicate entries, they decided only to store the lemma forms in the dictionary and to use a stemming algorithm to reduce words to their uninflected form. For example, the GI dictionary does not contain an entry for the word ‘loved’ (a predominantly positive word), but does contain an entry for the word ‘love’, which would be discovered as part of the stemming routine. The stemming routine itself is a multi-stage set of stemming rules designed to execute at runtime. At each stage of the stemming process a lookup is performed in the dictionary. If the word is found during lookup, then execution of the stemming algorithm is paused. If not, then the stemming algorithm proceeds to the next stage in the process, and repeats this process until either a positive lookup is found or the stemming algorithm ends. Pseudocode for the stemming algorithm is given in Appendix B.

The GI’s stemming algorithm naturally introduces a small runtime overhead. However, using the lemma form in conjunction with a stemming algorithm reduces the memory required to store the dictionary, both in terms of its primary and secondary storage. One final but very important point about the stemming algorithm is its ability to add specific ‘marker tags’ as part of the stemming routine, which are later used for word disambiguation (to be discussed). In the example ‘loved’, for instance, the stemming algorithm would not only reduce this inflected form to the uninflected form

‘love’, but would also assign the suffix tag ED to the word at runtime. Similar suffix tags are applied upon removal of other suffixes, such as the S tag upon removal of an -s/-es suffix, or ING tag upon removal of an -ing suffix. Note that a given word could involve the removal of multiple suffixes, during which time all applicable suffix tags would be applied to the word.

It is important to emphasize that each entry in the GI dictionary actually corresponds to a particular word sense rather than a unique word type. For instance, the word ‘charge’ appears 9 times in the dictionary, with each sense being tagged with a unique set of semantic categories. As a result, the GI is equipped to cope with polysemy and it is this issue to which I now turn the reader’s attention. As a running example, I use the word ‘kind’, which has five entries in the dictionary. Among its various senses are the meanings of ‘benevolent’, ‘species’/‘type’, idiomatic forms (‘kind of’, ‘in kind’) and the root form of the adverb ‘kindly’. Upon encountering the word ‘kind’ in a given text, one needs to decide which sense should be applied. There are a number of possible solutions to this problem. Firstly, the system could simply ignore the sense definitions and apply all content categories for all possible senses. Thus when encountering the word ‘kind’ all content categories for all of the five senses would be applicable. However, this approach introduces statistical errors, by both incorrectly identifying non-applicable content categories and by failing to identify applicable categories. A second approach would be to avoid polysemous words altogether, and simply apply none of the content categories when encountering an ambiguous word. However, this approach also introduces statistical errors by failing to identify an applicable content category when one should be present. This approach is especially serious given that frequent words in English tend to have a higher number of meanings (Crystal (2010)). Ignoring polysemy altogether would thus mean disregarding from the analysis the most frequent words in the text. A third approach would be to always apply the ‘dominant’ (most frequent) sense of the word. While this approach minimizes the probability of statistical errors, it still fails to adequately deal with the issue of polysemy, and ignores the rich division of word senses that make up the GI dictionary. A final approach is to attempt some sort of word sense disambiguation, and it is this decision which the creators of the GI system ultimately took.

The GI system was one of the first computational attempts at explicitly addressing the issue of word sense disambiguation. The task - which came to be known as the

disambiguation project (Stone (1969)) - involved the construction of a set of linguistic disambiguation rules over a period of 7 years. Using a team of over 30 people the researchers (Kelly & Stone (1975)) carefully examined the context of the most frequent ambiguous words in a large textual corpus, and ultimately derived a list of linguistic disambiguation rules in a form that could be processed by a machine. Each ambiguous entry in the GI dictionary has a set of associated if-then rules that instruct the system to examine a specified context window, and look for the presence (or absence) of a particular set of words or tags (which could be either semantic content categories, or the stemming (and other syntactic) marker tags previously described). Each test has success or failure branches, which instruct the system what to do upon the outcome of the test. Possible outcomes include applying a specific sense number and terminating the rule processing, transferring to another rule for that particular word, or transferring to another dictionary entry and executing the rules associated with that entry. The basic structure of this contextual test (together with examples) are given in Appendix A.

In addition to handling contextual disambiguation rules, the GI disambiguation rules also contains two other rule types. The first is a SUPV rule: a binary-branching rule to determine whether the current word is a verb or non-verb. The SUPV routine involves a list of 9 conditional rules that examine the context around the word to check for the presence of certain marker tags, including syntactic marker tags such as DET (determiner), PREP (preposition) and PRON (pronouns), and suffix tags such as ED, LY and S, which have been applied to words as part of the stemming, semantic tagging and disambiguation processes. Upon finding these tags, the system takes the verb branch; otherwise it takes the non-verb branch. The outcome branches instruct the system to either apply a specific word sense, transfer control to another rule, or transfer to another dictionary entry.

The other GI disambiguation rule type is the GOTO rule, which allows the system to transfer control to another dictionary entry. A common use case is for past-tense irregular verb forms. As content categories for irregular verb forms are likely to be the same as those of infinitive forms, the GI rule set simply lists GOTO rules for any irregular verbs which instruct the system to transfer control to the infinitive entry in the dictionary. For instance, the GI rule set contains a GOTO rule for the past tense verb ‘kept’, which instructs the system to transfer control to the relevant sense of the

word for ‘keep’. This allows the system to assign any content categories specific to ‘kept’ (such as a past-tense markers), before transferring control to the ‘correct’ (most likely) sense for ‘keep’ and assigning any content categories specific to the infinitive form. Another common use case is for certain spelling variants, such as non-American spellings. For instance, the GI rule set contains a GOTO rule for the British spelling of the word ‘aluminium’, which instructs the system to transfer to the American spelling ‘aluminum’ in the GI dictionary. These GOTO rules therefore both reduce duplication in the dictionary, as well as providing disambiguation support.

The GI disambiguation rules are evaluated by the authors in Kelly & Stone (1975)⁷, and also reported upon briefly in Stone (1969). The authors used an evaluation corpus of 185,000 tokens, comprising a subset of the Brown corpus (Francis & Kucera (1967)) and a variety of texts used in various General Inquirer studies. The procedure for evaluation involved using the GI disambiguation rules to label the corpus, and then having 5 human annotators manually check the accuracy of ambiguous terms occurring 10 or more times. This resulted in the evaluators checking a total of 64,253 ambiguous word tokens, comprising 671 unique word types. Of the 64,253 tokens, the authors report that 59,176 instances were tagged accurately by the GI disambiguation rules, yielding a precision score of just over 92%. On a type-by-type basis (weighting entries equally regardless of frequency), the authors report an accuracy of just under 90%. Kelly & Stone (1975) acknowledge that a potential weakness with their evaluation was that it is assumed the GI-annotated text was correct unless identified as an error by a human evaluator. This means that any human oversight would have favoured the GI rules - a criticism that is compounded by the fact that the authors do not report upon the amount of human error that was encountered during the evaluation. Another potential criticism is that it is not clear how many of these 671 types correspond to evaluative terms, and it may be the case that the precision degrades (or improves) if their method of evaluation were applied solely to evaluative terms. However, despite these criticisms, this impressive precision supports the efficacy of the GI disambiguation rules, and suggests that the rules are a worthwhile addition to this thesis.

The GI dictionary has been through four major revisions, during which time new

⁷The evaluation of the disambiguation procedures occupies a whole chapter in Kelly & Stone (1975), and the interested reader is invited to consider this work for a more detailed overview of the evaluation process.

words and categories were added to the dictionary and existing categories were re-structured. Following their most recent revision, the authors in Dunphy et al. (1974) conducted a systematic evaluation of the dictionary’s contents which consisted of three primary steps: an author validation step, an external validation step, and a predictive validation step. During the author validation step, the three authors in Dunphy et al. (1974) manually inspected ‘all entries for each category to test their ‘goodness of fit’ with the definition of that category’. This examination, they claimed, revealed many words which were inappropriately allocated to a particular category, and so were subsequently reallocated elsewhere⁸. In the external validation step, the authors solicited the assistance of 14 external evaluators at the University of New South Wales, Australia. The evaluators were split into 5 groups, and each group was instructed to systematically work through each of the entries within their assigned categories, ensuring that the words assigned to a particular category were consistent with that category’s definition. Any words that were not consistent with the rest of the words in a given category were set aside for weekly discussion with the supervising committee, until a consensus had been reached for that word. In the predictive validation stage the authors constructed a corpus of texts on particular topics (e.g. religious, political, military, economic texts), and measured the frequency counts of GI categories in those texts. They reasoned that if the GI categories were valid, then ‘we should be able to use them to distinguish between texts of similar known characteristics’. If a particular category was higher than expected for a given text type (e.g. if the GI’s economic tag was particularly high in non-economic texts) then they examined the words that were inflating the frequencies of this category, and reallocated them to alternative categories if appropriate.

In summary, the GI dictionary consists of almost 12,000 word senses, each tagged with up to 183 semantic categories derived from research in social psychology. Each entry in the GI system represents a particular word sense, and a set of linguistic rules are used to disambiguate words and derive the correct sense at runtime. A stemming algorithm is also used to reduced inflected forms to their dictionary form. Of particular interest for sentiment analysis studies is the tagging of words according to Osgood et al.’s evaluative affect dimension - namely the GI categories of positive and negative.

⁸Unfortunately they do not give any examples of this in Dunphy et al. (1974).

However, two GI categories - vice and virtue - comprise a subset of the negative and positive evaluative terms. This thesis is concerned with this subset of vice terms, and seeks to question whether these terms can be used in place of the more general set of negative terms to derive a proxy for investor sentiment. In the remainder of this chapter I turn to related work on sentiment analysis, discussing other methods for deriving a sentiment proxy and related work that explores the relationship between sentiment and financial markets.

2.3 Construction of Sentiment Proxies

In this section the question of how to compute a sentiment proxy is discussed. Specifically, this section is interested in how to project qualitative textual data onto a quantitative sentiment variable that can be incorporated into statistical analysis. The existing literature on constructing text-based sentiment proxies can typically be divided into two approaches: dictionary-based methods (which use a pre-defined list of sentiment-bearing lexical items) and machine learning methods (which attempt to induce a set of features from the data that are indicative of sentiment). One such example of a dictionary-based approach was presented in the previous section. Of particular interest for sentiment analysis studies are the evaluative categories of positive and negative. By analysing the frequency distribution of positive and negative categories in a text, one can make inferences about the levels of evaluative language contained therein. In the case of sentiment analysis, one can use a suitable set of financial texts to make inferences about investors' opinions on a wide range of subject matter.

There are several sentiment dictionaries that have been used in research on sentiment analysis, and this section aims to cover some of the main dictionaries used in the literature. Additionally, I also discuss machine learning approaches which attempt to use statistical learning algorithms to learn words, phrases and other features associated with movements of key financial market variables.

2.3.1 Dictionary-Based Approaches

Dictionary-based methods of constructing a sentiment proxy have been extremely popular in the literature on sentiment and finance. At its core, a dictionary-based approach involves computing frequency counts of sentiment-bearing words and phrases labelled

in a pre-compiled sentiment dictionary. Typically, the categories of negative or positive are used or some net combination thereof. However more-fine grained emotional categories can also be used (e.g. anxiety, worry, fear (Gilbert & Karahalios (2010))), depending upon the research question. The main premise behind a dictionary-based approach is to take a document as input and, for each word in the document, perform a lookup in the sentiment dictionary to see whether the word is present. A running-total of the number of occurrences of each sentiment category is kept for each document in the corpus. The output of this method is a frequency matrix, in which columns represent sentiment categories (e.g. positive, negative) and rows represent individual documents. Attempts can also be made to weight the terms according to the degree of association with each sentiment category. As the raw frequency is a poor indicator of the sentiment in each document (longer documents naturally contain more sentiment-bearing words and phrases), the frequency is usually normalized in some fashion, such as by the total number of words in each document.

According to Loughran & McDonald (2015) there are four major sentiment dictionaries that have been widely used in the text analysis literature in accounting and finance⁹. These are:

- Diction (Diction Software (2017))
- Henry’s sentiment dictionary (Henry (2008))
- Loughran and McDonald’s sentiment dictionary (Loughran & McDonald (2011))
- General Inquirer IV-4 dictionary (Stone et al. (1966), Kelly & Stone (1975))

The Diction dictionary (DICTION) is packaged as part of the Diction content analysis software, and has been used in a variety of text analysis studies¹⁰. In the accounting and finance domain specifically, the dictionary has been used to explore market responses to earnings press releases (Demers & Vega (2010), Davis et al. (2012)), annual reports (Wisniewski & Sina Yekini (2015), Davis & Tama-Sweet (2012)), and in exploring the impact of initial public offering (IPO) prospectuses (Ferris et al. (2013)). Like

⁹I am aware of the existence of several other sentiment dictionaries such as the Linguistic Inquiry and Word Count (LIWC), Multi-Perspective Question Answering (MPQA) subjectivity lexicon, SentiWordNet affect and Bing Liu’s opinion lexicon, but these are not typically used for performing sentiment analysis in the accounting and finance domain.

¹⁰A range of example studies can be found on the authors’ website: www.dictionsoftware.com, Last Accessed: 21 August 2017

the General Inquirer dictionary, DICTION is a general language dictionary, designed to be used with any types of text. This contrasts with the Loughran and McDonald (L&M) and Henry dictionaries, which were specifically designed for use in financial texts. In its current form (version 7.0), DICTION comprises 5 main semantic categories (Activity, Optimism, Certainty, Realism and Commonality), and 35 subcategories (e.g. Tenacity, Blame, Ambivalence, Motion, Communication). The Optimism category (a combination of the three subcategories Praise, Satisfaction and Inspiration) is typically used as a proxy for positive investor sentiment. Although the dictionary lacks an equivalent Pessimism category to proxy for negative investor sentiment, researchers have typically combined the sub-categories of Blame, Hardship and Denial (Loughran & McDonald (2015)). Despite its widespread usage Loughran & McDonald (2015) claim that DICTION is inferior to their own dictionary in the categorization of financial or business texts. However, the L&M dictionary was constructed from the most frequently occurring words in a corpus of Form 10-Ks¹¹, whereas the DICTION was hand-constructed as a general language dictionary designed to work with a variety of texts. Comparing the two dictionaries' coverage in this manner is equivalent to comparing in-sample and out-of-sample statistical performance. Furthermore, while Loughran & McDonald (2015) report a small anti-correlation (-0.023) between the positive word lists in DICTION and their own dictionary, the correlation between the negative word lists is relatively strong (0.688). Indeed in subsequent statistical analysis where the authors in Loughran & McDonald (2015) use both dictionaries in a model to forecast future volatility, it is revealed that both dictionaries exhibit similar conclusions using their respective negative word lists (though opposite findings when using the positive lists). However, this is not to say that Loughran & McDonald's criticisms are without merit. Their finding that a number of words in DICTION are misclassified in financial or business contexts is corroborated by other studies such as Davis & Tama-Sweet (2012) and Ferris et al. (2013). The implications of this are that a number of entries in the DICTION dictionary could introduce statistical errors, which could overestimate the impact of sentiment and jeopardise any statistical claims.

¹¹Form 10-Ks are annual financial performance reports that all US companies are required to file with the Securities and Exchange Commission. These reports provide a mixture of qualitative and quantitative information, reflecting on both past and estimated future performance. Researchers in sentiment analysis tend to be interested in the Management Discussion and Analysis sections, where the company's management team qualitatively discusses the company's operations.

Henry’s (2008) sentiment dictionary was originally constructed for analysing the impact of sentiment in earnings press releases. Using a corpus of 1,366 press releases for firms in the telecommunications and IT industries, the author created a list of the most frequently occurring negative and positive terms. An aggregate measure of tone ($\frac{\text{positives}-\text{negatives}}{\text{positives}+\text{negatives}}$) was then used to examine the impact of sentiment on abnormal returns. Since then the dictionary has been used in a number of other studies, including Sadique et al. (2008) where the dictionary was used to measure the effect of tone on stock returns and volatility, and Price et al. (2012) where it was used to measure the impact of sentiment in earnings conference calls¹². A number of studies have also combined Henry’s dictionary with either the LM/GI/DICTION dictionaries in order to improve coverage (see, for instance Huang et al. (2014), Rogers et al. (2011), Davis et al. (2015)). Henry’s dictionary is by far the smallest of all four dictionaries (see Table 2.3 for the size of each dictionary), and it is clear that the author is attempting to maximize precision in the target domain at the expense of recall in other domains. Henry’s dictionary has also undergone a number of criticisms. In Loughran & McDonald (2015), for instance, the authors suggest that the dictionary has poor coverage of Form 10-Ks relative to their own dictionary. Specifically, Loughran & McDonald (2015) claim that highly frequent negative words (such as ‘loss(es)’, ‘claims’, ‘impairment’, ‘against’ and ‘adverse’) are absent from the list, and that the dictionary suffers from poor coverage. However, the L&M dictionary was constructed specifically from Form 10-Ks, whereas Henry’s dictionary was constructed from earnings press releases. As a result it does not make sense to compare the coverage of a dictionary constructed from earnings press releases in a corpus of Form 10-K texts. However, Loughran & McDonald’s (2015) criticism does raise an interesting point. Despite the dictionary’s reported quality in Price et al. (2012), perhaps the dictionary does not perform so well outside of texts about earnings press releases.

The Loughran and McDonald (L&M) sentiment dictionary is one of the most widely used dictionaries in the literature on sentiment analysis (the other being the General Inquirer). The L&M dictionary was originally constructed by examining the frequency

¹²Earnings conference calls are teleconference or webcasts in which company management broadcast information on their earnings to investors, analysts and members of the public. These calls allow the management of a company to draw attention to recent successes, and to allay fears during adverse times (Investopedia (2005)). They are typically conducted at the end of each quarter, and are usually accompanied by an earnings press release (ibid.)

of words in a sample of 50,115 Form 10-Ks filed with the Securities and Exchange Commission between 1994 and 2008. To create the list of positive and negative terms, the authors hand-analysed all words occurring in at least 5% of the documents in the corpus, selecting positive and negative words which they believed were unequivocally positive or negative in a financial reporting context¹³. The authors subsequently used the list of negative words to explore the impact on the market around the time of 10-K filings. They observed that the negative terms had a significant impact on excess returns around the filing date, and find that their wordlist also had an impact on trading volume, unexpected earnings and market volatility. A number of other studies have also used the L&M dictionary for sentiment analysis in financial markets. For instance, García (2013), Ferguson et al. (2015), Ahmad et al. (2016) and Jegadeesh & Wu (2013) used the dictionary to examine the effect of news sentiment on financial returns. Huang et al. (2014) used the dictionary to examine tone in earnings press releases. While the LM dictionary appears to boast impressive coverage of negative terms relative to the other dictionaries (see Table 2.3), it should be noted that the reported counts in Table 2.3 comprise both lemma and inflected forms¹⁴. Thus the reported number of words in the dictionary overestimates the true number of unique word types.

Of all four sentiment dictionaries perhaps the most widely used is the General Inquirer dictionary (Stone et al. (1966)). I have already discussed the construction and composition of the GI dictionary in the previous section, and will not repeat this discussion here. However, I take this opportunity to now consider some of the studies that have used the GI dictionary, together with some of its advantages relative to the other aforementioned sentiment dictionaries. The GI dictionary has been used to examine the effect of sentiment in a number of sources, including news (Tetlock (2007), Tetlock et al. (2008), Engelberg (2008), Cook & Ahmad (2015), Feldman et al. (2010), Kothari et al. (2009), Twedt & Rees (2012), Heston & Sinha (2016)), message board postings (Das & Chen (2007)), IPO prospectuses (Hanley & Hoberg (2010)), earnings conference calls (Doran et al. (2012)), and many others (Huang et al. (2014), Ferguson et al. (2015), Chen et al. (2014)). Like the DICTION dictionary (and in contrast to

¹³In its current form the dictionary also contains categories for uncertainty, litigious, constraining, superfluous, interesting, modal verbs and irregular verbs.

¹⁴For instance the LM dictionary contains 6 entries for the word ‘abandon’, including ‘abandon’, ‘abandoned’, ‘abandoning’, ‘abandonment’, ‘abandonments’ and ‘abandons’.

the LM and Henry dictionaries) it is a general language dictionary and is not confined to a particular domain. The GI's primary advantage over the other three dictionaries is its handling of polysemous terms. Unlike entries in the DICTION, Henry and LM dictionaries, the entries in the GI dictionary are based on individual word senses and can thus handle the multitude of meanings a given word can have. A set of linguistic disambiguation rules (discussed in Section 2.2.4) can be used in conjunction with the dictionary in order to identify the most appropriate sense in a given context. A second advantage of the GI dictionary is that its composition process was far more rigorous than the composition processes of the other dictionaries. While the other dictionaries were constructed by one or two authors at most, the GI dictionary was constructed by a whole team of social psychologists and was subsequently revised over a period of many years. Maximizing inter-annotator agreement was paramount in the construction of the dictionary in order to minimize subjectivity of the dictionary's creators. Indeed the fact that the GI dictionary is still being used half a century after its original construction is testament to its quality. A third advantage of the GI dictionary over the other sentiment dictionaries is its coverage. Table 2.3 shows that the GI dictionary has more positive entries than any of the other major sentiment dictionaries, and the second highest number of negative entries. However these numbers vastly under-report the true number of sentiment bearing words in the GI dictionary. The entries in the GI all appear in their canonical form, and the GI dictionary is designed to be used in conjunction with a stemming algorithm to reduce inflected forms to their lemma. Thus in reality the GI dictionary covers a much larger number of words than reported, implying that the GI coverage is larger than any of the other major sentiment dictionaries discussed here. Another motivating factor for using the GI dictionary is that the dictionary contains the ethical categories of vice and virtue, which I use in this thesis to explore the hypothesis that sentiment is ethically motivated. Finally, the GI dictionary's list of negative and (to a lesser extent) positive words have been repeatedly shown to have an impact on financial markets, providing further motivation for its use.

Despite its success in sentiment studies, there have also been some concerns over the GI's quality. For instance, Loughran & McDonald (2011) have criticised the General Inquirer's coverage in the domain of annual reports. Like their critique of Henry's and DICTION's dictionary, they argue that the General Inquirer has poor coverage of Form 10-Ks and that a number of domain-specific words are misclassified in the domain of

business and finance. Examples of reportedly misclassified words include ‘tax’, ‘capital’, ‘expense’, ‘cancer’ and ‘mine’. However, the authors in Loughran & McDonald (2011) ignored the fact that the GI has routines in place to deal with polysemy. Whereas entries in the LM dictionary denote words, the entries in the GI dictionary are individual word senses, designed to be used in conjunction with a set of disambiguation rules. Consider the word ‘capital’, for instance, which the authors claim is misclassified as negative in financial texts. The GI dictionary actually contains four senses for the word ‘capital’, only one of which is tagged as negative . When used in conjunction with the disambiguation rules, the only time that ‘capital’ is considered a negative word is if it occurs as part of the phrases ‘capital punishment’, ‘capital crime’, or ‘capital charge’ which have nothing to do with the domain of finance. Likewise, the word ‘tax’ also has four senses, only one of which is tagged as negative in the GI dictionary. However, this is not to say that the authors criticisms are without merit. The word ‘cancer’ in their exemplar list is unequivocally classified as negative in the GI dictionary. In certain domains, such as the health care domain, the word ‘cancer’ does not carry evaluative connotations, and thus should not be classified as negative in these domains. The fact that domain-specific lexical items can unintentionally inflate the sentiment proxy is a valid concern. To overcome this issue I introduce the notion of a domain-specific ‘override’ dictionary, containing frequent words and phrases from the domain of study. This domain override dictionary, which will be discussed in greater depth in the next chapter, instructs the system to ignore the evaluative meaning of words in the target domain.

Table 2.3: Frequency of negative and positive words in the Diction, Henry, Loughran & McDonald and General Inquirer dictionaries.

Dictionary	Negative	Positive
Diction	920	686
Henry	85	105
Loughran & McDonald	2,329	354
General Inquirer	2,291	1,915

In this section I have looked at four dictionaries that have been used extensively in the text analysis literature on finance and accounting. These dictionaries have been used to construct a sentiment proxy, by mapping words and phrases onto a sentiment variable. There are advantages and disadvantages with all four dictionaries, however constructing a sentiment dictionary with 100% inter-annotator agreement is an impossible task. Additionally, attempting to use a dictionary outside of the domain for which it was constructed could lead to lower precision and coverage in the target domain. In the next section I briefly discuss some of the machine learning methods that have been used to construct sentiment proxies.

2.3.2 Machine Learning Approaches

A number of studies have used machine learning methods to project textual data onto a quantitative sentiment proxy. These studies tend to use supervised learning techniques (Nassirtoussi et al. (2014)), particularly classification methods such as Naïve Bayes, support vector machines, decision trees and conditional random fields¹⁵. In supervised learning, the main objective of the learning algorithm is to infer a statistical function from labelled examples. These labelled examples are typically referred to as the ‘training set’, to distinguish it from a separate set of labelled examples (the ‘test set’) which are used to evaluate the precision of the classifier.

Many of the machine learning approaches to sentiment analysis adopt a bag-of-words methodology (e.g. Wüthrich et al. (1998), Mittermayer (2004), Kogan et al. (2009), Huang et al. (2014)), in which a text is represented as unordered sequence of words. Each individual word would be considered a ‘feature’ in a high dimensional space, and a learning algorithm is then typically used to learn the features (here words) that are associated with movements of particular market variables such as stock prices¹⁶. The extracted words may or may not be linguistically evaluative, in contrast to the dictionary approach which is usually constructed with some psychological theory in mind. Nevertheless, a number of previous studies have attempted to learn key words and

¹⁵Unsupervised techniques have also been used to a lesser degree. In unsupervised learning the goal is to attempt to learn infer a function without explicitly labelled training data. However, the present task - attempting to classify documents as positive or negative - is better suited to supervised techniques which generally give a higher degree of accuracy.

¹⁶In theory any set of countable features could be used, however a common approach in text analysis is to learn the probabilities of particular words and phrases associated with certain market movements.

phrases associated with market movements. One of the earliest studies was that of Antweiler & Frank (2004), who used a Naïve Bayes classifier to examine 1.5 million posts downloaded from the Yahoo! Finance and RagingBull.com message boards. The authors manually classified 1,000 of these messages as either a ‘buy’, ‘hold’ or ‘sell’ message, and then used a Naïve Bayes algorithm to learn the probabilities of each word being associated with each class. After using these probabilities to classify the remaining texts in their corpus, the authors developed two sentiment indices: a bullishness index (an aggregated and scaled measure that indicated the difference between the number of buy and sell posts) and an agreement index (indicating the extent to which the message posters agreed or disagreed). They subsequently used these measures to explore their effect on financial returns, trading volume, and volatility. Their main findings were that while neither the bullishness nor disagreement index had a statistically significant impact on forecasting future returns, both had a significant impact on future volatility and future trading volume. In addition, they found that the number of message posts had a significant impact on forecasting returns, volatility and trading volume.

Li (2010b) also used a Naïve Bayes classifier, this time in the context of examining forward looking statements in the Management, Discussion & Analysis Section of US annual (10-K) and quarterly (10-Q) reports. Unlike Antweiler & Frank (2004), who adopted a document-level approach, Li’s (2010b) approach worked at the sentence level. Using a random sample of 30,000 sentences as a training corpus, the author annotated each sentence as being either positive or negative, together with the topic of each sentence (e.g. profitability operations, liquidity, etc.). The author then used the Naïve Bayes algorithm to categorize the sentiment and tone of 13 million sentences taken from 140,000 10-Q and 10-K filings. He subsequently found that firms with more positive forward-looking statements tended to be those with better current performance, lower accruals, smaller size and less return volatility.

The Naïve Bayes classifier has been shown to be useful in deriving a sentiment proxy, however it is not the only learning algorithm that has been used. Support vector machines (SVM) have also commonly been used for deriving a sentiment proxy. For instance, Mittermayer (2004) performed SVM classification of press releases to classify ‘buy’, ‘hold’ and ‘sell’ recommendations. The authors report little difference between their system and a random trader. Schumaker & Chen (2009) also use SVM

classification in the analysis of news articles about companies in the S&P 500 index over a five-week period. The authors compared three separate features - individual bag-of-words, noun phrases and named entities - finding that named entities were more useful than a baseline bag-of-words model in the prediction of equity returns. Fung et al. (2003) used a weighted bag-of-words approach in which words are weighted by their tf-idf weighting scheme. Using a corpus of 350,000 news articles the authors find that mining multiple time series concurrently substantially improved their classification algorithm.

In addition to Naïve Bayes and SVM classification/regression, authors have also generated sentiment proxies through the use of genetic algorithms (Thomas & Sycara (2000)), fuzzy neural networks (Bollen et al. (2011)), conditional random fields (Chen et al. (2014)), probabilistic language models (Lavrenko et al. (2000)), multiple kernel learning (Shynkevich et al. (2016)) and many others. Attempts have also been made to use ensemble learning (Das & Chen (2007), Wüthrich et al. (1998)), whereby multiple algorithms are combined.

2.3.3 Advantages and Disadvantages

In this thesis I adopt a dictionary-based approach to constructing a sentiment proxy. There are several advantages of dictionary-based methods over machine learning approaches, and I briefly discuss these here. Firstly, when using a pre-defined sentiment dictionary the researcher does not require explicitly labelled training data, which can be costly, time-consuming and introduce subjectivity on behalf of the annotator. Secondly, while a dictionary of sentiment bearing terms may be linguistically or theoretically grounded, training features selected for a particular machine learning task might not be meaningful, thereby potentially reducing the generalizability of the fitted model to other data sets. Dictionary-based approaches generally tend to use ‘core’ lists of sentiment bearing words and phrases terms, which may generalize better than fitted machine learning models when applied to other domains or data sets. Finally, dictionary-based approaches are generally easier to replicate than machine learning approaches, which often have a number of tuning parameters that are not typically specified in the literature. This is an obvious benefit for other researchers, and in Chapter 4 I replicate and extend the results of a major existing study, using a dictionary-based approach.

However, there are also several disadvantages of using dictionaries to form a sentiment proxy. Given the time and manual effort involved in their construction, they can sometimes be limited in size, whereas machine learning methods can be very scalable. This problem could be compounded due to changes in language, which introduce new words into the lexicon (neologisms), remove old words (archaisms), or change the meaning of existing words¹⁷. Dictionaries constructed for a particular field may also not scale well to new fields leading to losses in both precision and recall. While dictionary-based approaches are generally limited to analysing positive and negative words and phrases, machine learning approaches are capable of working with features above the word-level (e.g. sentence length or document complexity). Finally, while the use of a pre-defined sentiment dictionary reduces subjectivity on behalf of the researcher using it, it could introduce a layer of subjectivity on behalf of the person constructing it.

2.4 Related Work in Sentiment Analysis

So far I have discussed the traditional theories underpinning asset pricing models in finance. I have shown that the traditional efficient market hypothesis, which dictates that markets are informationally efficient, is often unable to describe many market anomalies. Behavioural finance was subsequently presented as an alternative paradigm, and showed how investors can be impacted by emotional, psychological and behavioural factors that cause them to act ‘irrationally’, by valuing asset prices above or below their ‘true’ (i.e. fundamental) value. Several sentiment indicators have also been discussed which are often used to assess the value of a particular asset. In this section I present an alternative view of valuing investor sentiment, which attempts to automatically construct a sentiment proxy by analysing textual information. While I acknowledge that sentiment analysis has played a role in areas other than finance (for example in mining sentiment in reviews of products and services, determining the effectiveness of marketing feedback, etc.) this section will focus on studies that refer to sentiment in finance. The interested reader is invited to consider Liu (2012), Liu & Zhang (2012) and Liu (2015) for a synthesis of research in these other areas.

One of the earliest studies in exploring the impact between news and financial

¹⁷In slang vernacular, for instance, words can gradually change their polarities. The words ‘cool’, ‘wicked’ and ‘sick’ can have positive connotations in certain contexts.

markets was that of Niederhoffer (1971). In this study, the author examined the impact of major world events (defined by headlines in the New York Times) on key stock indices such as the Dow Jones Industrial Average. The author concluded that on days following major news events there is a significant change to the DJIA. Consequently, this research demonstrated the impact that information flow has on financial markets. Since this point there have been many studies on sentiment analysis, and this section explores some of these studies.

There have been several recent surveys on the state-of-the-art literature in text analysis in finance and accounting (Loughran & McDonald (2016), Kearney & Liu (2014), Nassirtoussi et al. (2014), Li (2010a), Das (2014)). The aim here is not simply to re-hash these surveys, but rather to pick out the pertinent studies related specifically to text analysis. In describing related work in the sentiment and finance/accounting literature I separate my review into three major sections. In the first section I group studies by their textual information source, describing the different types of news and non-news texts used in this domain. In the second section I describe some of the key statistical methods that have been used to explore the relationship between sentiment and economic data. This second section provides a thorough grounding of some of the statistical methods used in this thesis, and situates them within the body of existing work. Finally, in the third section I describe some of the key market variables that have been explored in the literature.

2.5 Information Sources

It is often believed that news and other textual information play a fundamental role in the determination of asset prices. While textual information may contain facts that have yet to be incorporated into an asset's price, texts may also reveal the general sentiment of the market - that is, whether investors are generally feeling bullish (optimistic) or bearish (pessimistic) at an aggregate level. This sentiment can subsequently be used to gauge whether investors perceive a particular instrument to be overvalued or undervalued, relative to its fundamental value.

With the advent of the Internet, the digitalisation of texts, and the massive increase in user-generated content, investors and traders have access to more information than ever before. As the number and variety of textual sources continues to grow at an

overwhelming rate, new systems are needed to assist in extracting and summarizing this information into a format that can be more easily digested by the reader. In this section I review some of the major sentiment analysis studies, grouping them according to their textual information source¹⁸. In discussing the various information sources, I follow a similar approach to Nassirtoussi et al. (2014), grouping texts into three primary groupings: news texts, reports and disclosures and social media texts:

1. News: Including newspapers, radio, television, newswires, fax and the web. Generally these texts are from more reputable sources.
2. Reports and Disclosures: The primary source material used by reporters to write the news. For instance, company annual and quarterly financial reports, IPO prospectuses, court documentation and other corporate or governmental disclosures.
3. Social Media: Such as message boards (forums), blogs, microblogs and other social media. Generally these are of a less reputable quality, as the barrier to publication is much lower.

Table 2.4 provides a summary of pertinent literature grouped according to text type. An overview of each source, with a discussion of key studies, is provided in the following subsections.

¹⁸Although there have been a number of studies on the ‘readability’ of texts (which explore the impact of a document’s readability in determining asset prices), these studies are beyond the scope of this thesis. The interested reader is invited to consider Loughran & McDonald (2016) for a synthesis of some of the major readability literature.

Table 2.4: Summary of pertinent literature by text type. Op-Ed News refers to opinion columns and/or editorials. The No. Texts column denotes the size of the textual corpus in each study. The abbreviation DJNS refers to the Dow Jones News Service, and 10-K and 10-Q refer to annual and quarterly financial reports that US companies are obligated to file with the Securities and Exchange Commission (SEC).

Panel A: News				
Text Type	Text Source	Period	No. Texts	Ref.
Op-Ed News	Wall Street Journal	1984 - 1999	3,709	Tetlock (2007)
Op-Ed News	New York Times	1905 - 2005	55,307	García (2013)
General News	DJNS, Wall Street Journal	1980 - 2004	350,000	Tetlock et al. (2008)
General News	LexisNexis News	2001 - 2010	5.5 million	Ahmad et al. (2016)
General News	4 UK Newspapers	1981 - 2010	264,647	Ferguson et al. (2015)
Business News	DJNS	1980 - 1999	500,000	Chan (2003)
Business News	DJNS	1999 - 2005	51,207	Engelberg (2008)
Business News	Biz Yahoo!	1999 - 2000	38,469	Lavrenko et al. (2000)
Business News	Reuters NewsScope	2007 - 2008	29,497	Groß-Klußmann & Hautsch (2011)
Newswire	PR Newswire	1998 - 2003	23,017	Davis et al. (2012)
Newswire	PR Newswire, Business Wire	2009 - 2014	51,000	Shynkevich et al. (2016)
Newswire	PR Newswire	2002	6,602	Mittermayer (2004)
Newswire	PR Newswire, Business Wire	1997 - 2007	18,436	Huang et al. (2014)
Newswire	PR Newswire, Business Wire	1998 - 2002	1,366	Henry (2008)
Newswire	PR Newswire	1998 - 2006	27,705	Demers & Vega (2010)

Panel B: Reports and Disclosures				
Text Type	Text Source	Period	No. Texts	Ref.
Company Reports	10-K Reports	1994 - 2008	50,115	Loughran & McDonald (2011)
Company Reports	10-K and 10-Qs	1993 - 2007	153,988	Feldman et al. (2010)
Company Reports	10-K and 10-Qs	1995 - 2002	45,860	Jegadeesh & Wu (2013)
Company Reports	UK annual reports	2006 - 2012	1,262	Wisniewski & Sina Yekini (2015)
Company Reports	IPO Prospect-uses	1996 - 2005	2,112	Hanley & Hoberg (2010)
Panel C: Social Media				
Text Type	Text Source	Period	No. Texts	Ref.
Social Media	Twitter	2008 - 2008	9.9 million	Bollen et al. (2011)
Social Media	Twitter	2009 - 2009	8,100 - 43,040 per day	Zhang et al. (2011)
Social Media	Twitter	2011 - 2011	5 million	Vu et al. (2012)
Social Media	Twitter	2012 - 2013	624,782	Si et al. (2013a)
Social Media	StockTwits.com	2010 - 2010	208,278	Oh & Sheng (2011)
Blogs	LiveJournal	2008	20 million	Gilbert & Karahalios (2010)
Message Boards	RagingBull.com	1999 - 2000	181,633	Tumarkin & Whitelaw (2001)
Message Boards	Yahoo! Finance, RagingBull.com	2000	1.5 million	Antweiler & Frank (2004)
Message Boards	Yahoo! Finance	2005 - 2010	32 million	Kim & Kim (2014)

2.5.1 News

One of the earliest and most widely cited studies on sentiment in finance is the work conducted by Tetlock (2007), who was one of the first authors to explore the impact of sentiment in news on financial markets. Using a daily opinion column from the Wall Street Journal, he discovered that a high proportion of negative sentiment predicted downward pressure on next-day financial returns of the DJIA index. As a proxy for negative sentiment, Tetlock used the (standardized) fraction of negative words in each article, calculated using the GI system (Stone et al. (1966), Kelly & Stone (1975)). The methods in Tetlock's study form the foundation of this thesis, and in Chapter 4, I replicate and extend some of his key results, using the system constructed in this thesis.

García (2013) also used daily opinion columns as a source of information for investor sentiment. Whereas Tetlock used a column from the WSJ over a 16-year period (1984 to 1999), García used two daily opinion columns from the New York Times over a much longer period (1905 to 2005). However, the structure of the textual sources used in both Tetlock's and García's studies was essentially the same: both opinion columns were a summary of US financial market activity on the preceding day, describing movements in key indices, key company movements, general market events and occasionally some speculation about what the market will do next. Unlike Tetlock, who used negative words from the GI dictionary to construct his sentiment proxy, García used negative words from the L&M dictionary (Loughran & McDonald (2011)). In addition to reaffirming Tetlock's results over a longer time period, García found that the effect of negative sentiment on next-day DJIA returns was stronger during recessionary periods than during non-recessionary periods. He attributed this to the fact that investors more closely follow the news during recessions. García also reported that the effect of sentiment on next-day DJIA returns was stronger after a non-trading day, such as on a Monday or the day after a statutory holiday. This result, he postulated, was due to the fact that investors had more time to read and digest the news during non-trading days.

The op-ed columns described in the preceding paragraphs represent a very specific subset of news. Given that these texts are, by their very nature, opinionated, a sceptical reader might conjecture that they are more likely to contain opinion-bearing words and phrases than might be found in more general news articles. Can, then, it be shown

that Tetlock and García’s conclusions about the impact of negative sentiment on future market performance stand up to scrutiny using a more general news source?

A number of authors have also attempted to explore the impact of sentiment in general news¹⁹ (Lavrenko et al. (2000), Tetlock et al. (2008), Ferguson et al. (2015), Chan (2003), Schumaker & Chen (2009), Groß-Klußmann & Hautsch (2011), Engelberg & Parsons (2011), Sinha (2016)). In Tetlock et al. (2008), for instance, the authors use news from the Dow Jones News Service (DJNS) and Wall Street Journal (WSJ) about S&P 500 firms. The authors find that an increase in negative words in these articles forecasts low future firm earnings. Lavrenko et al. (2000) explore the impact of general business news collected from Biz Yahoo! for 127 stocks during over a 5 month period. The authors use a probabilistic language model to learn words which are likely to precede a downward or upward trend in firm-level share prices. Groß-Klußmann & Hautsch (2011) and Sinha (2016) use news obtained from the Thompson Reuters NewsScope engine, where news stories are hand-annotated with relevancy scores for each news item. Their findings support the hypothesis of news influence on volatility and trading volume. Ahmad et al. (2016) build a very large corpus of 5.5 million articles on 20 large US firms over a 10 year period. Unlike previous studies, which have tended to focus on news from one or two sources, the authors used a variety of English language newspapers published via the LexisNexis news archive. The authors find that negative sentiment from their firm-level news corpus has a significant effect on next-day firm-level returns.

A number of studies have also investigated news distributed via newswire services. While historically newswire was syndicated via telegraph and was therefore more timely than the printed press, in today’s digital age the distinction between newspapers and newswire is more blurry. Both newspapers and newswire are considered to be secondary sources, however newswire texts are arguably more primary than their newspaper counterparts. Newswire services are typically used by companies and PR agencies to disseminate focussed and relevant stories to newspaper journalists, who use it as primary reference material for their stories²⁰. Consequently, newswire services offer ‘A

¹⁹By general news I mean news that appears across all columns in a paper, rather than confined to a specific type of news such as op-ed, together with web news in general. I do not cover radio or television news as the focus of this thesis is on analysing textual data.

²⁰According to Henry (2008) and Munz (2011), newswire services are also the main method for distributing earning’s press releases; the US Securities and Exchange Commission requires companies

vast amount of textual information (that) is available in real time' (Gonzalez (2011)). In the trading world, a number of several premium newswire vendors (e.g. Dow Jones Newswire Services, Thompson Reuters, Ravenpack) offer access to dedicated financial news and, as a result, 'Most traders have access to newswires at their desks' (Mitra & Mitra (2011)).

One of the earliest studies on sentiment in newswire texts was Chan (2003), who used a combination of newspapers and newswire headlines to examine the effect of sentiment on monthly returns. The author found that the Dow Jones newswire services offered more complete coverage across time and stocks than traditional newspaper services. Mittermayer (2004) developed a news categorization system, which took press releases from PR Newswire and made trading decisions by categorizing the articles into either 'Good News' (news which results in a rise in share price), 'Bad News' (news which precedes a fall in share price) or 'No movers' (news which does not result in a change in share price in either direction). Using intraday trading data the author found that the system significantly outperformed a random trader who randomly bought or sold stocks immediately after the publication of the press release. PR Newswire has indeed been used in a multitude of studies - in Davis et al. (2012) and Davis & Tama-Sweet (2012) the authors used it to analyse sentiment from earnings information, and in Luss & d'Aspremont (2015) the authors use PR Newswire to explore intraday price movements. Other authors (Huang et al. (2014) and Shynkevich et al. (2016)) combine texts distributed via PR Newswire and Business Wire, and it has been observed that these news sources combined account for more than 99% of all company press releases (Mittermayer (2004)). Other popular newswire sources include Ravenpack (Ho et al. (2013)) and Dow Jones Newswire Services (Boudoukh et al. (2013), Engelberg (2008), Chan (2003)) - premium newswire providers that offer specific financial news.

2.5.2 Corporate Disclosures

Several recent studies have utilised corporate or governmental disclosures as a source of investor sentiment. These materials typically comprise the primary-source material researched by journalists and other writers in order to produce mainstream news.

to file quarterly and annual earnings reports, and businesses typically release earnings press releases at the same time.

Principally, this information source encompasses government documents such as court documents, SEC documents and other government reports, and corporate documents such as quarterly and annual reports, letters to stakeholders, financial statements, and initial public offering (IPO) prospectuses. An excellent overview of the text analysis literature employing corporate documentation is given in Li (2010*a*).

The majority of literature on corporate disclosures has focussed on analysing sentiment in quarterly and annual reports (e.g. Jegadeesh & Wu (2013), Li (2006), Loughran & McDonald (2011), Kothari et al. (2009), Huang et al. (2014) and Feldman et al. (2010)). By law, the US Securities and Exchange Commission (SEC) requires US public companies to file periodic reports on the company's financial performance, which are subsequently disseminated to the public via the SEC website. Amongst other things, these reports declare risk factors facing the company, descriptions of legal proceedings, financial statements, and discussions of organizational hierarchy. Of particular interest in sentiment analysis studies is the Management Discussion & Analysis (MD&A) section, in which company management provides a qualitative interpretation of the company's financial performance over the previous year, and outlines forthcoming goals and projects for the next year. Although managers naturally face pressure to portray their companies in the best possible light, the SEC instructs them to do so in plain English. Companies also face severe legal penalties for misstatements. However, regardless of any perceived bias, the aforementioned studies have found these periodic reports to be a useful source of investor sentiment.

Another useful source of corporate disclosures information is initial public offering (IPO) prospectuses. Like quarterly and annual reports, IPO prospectuses are legal documents that companies are required to file with the SEC upon going public. In these prospectuses companies provide initial details about investment offerings that will be sold to the public. By analysing the information content of these documents, investors hope to ascertain whether an IPO will be overpriced or underpriced. In Jegadeesh & Wu (2013), for instance, the authors analyse a sample of 1,475 IPO prospectuses over the period 1995-2010. The authors observed that, on average, an increase in the negative terms in IPO prospectuses generally resulted in the IPO being underpriced. These results mirror similar results to Hanley & Hoberg (2010), who found that higher net tone (percentage of positive minus percentage of negative terms - determined by the GI dictionary), led to lower first-day returns, and smaller changes in the offer price

revision.

A number of authors have also analysed sentiment from earnings conference calls. For instance, Doran et al. (2012) analyse sentiment in transcripts from quarterly earnings conference calls for Real Estate Investment Trusts. Using the GI system, the authors found that tone (a combination of the categories negative, positive, active, passive, weak, strong, overstate and understate) had significant explanatory power for contemporaneous stock price prediction. Price et al. (2012) also examined the transcripts of quarterly earnings conference calls, and found that linguistic tone (calculated using the GI and Henry dictionaries) was significantly related to abnormal returns, abnormal trading volume, and post-earnings-announcement-drift. Davis et al. (2015) also focused on sentiment in earnings conference calls, but specifically analysed manager-specific optimism. The authors found that manager-specific tone in earnings conference calls had a statistically significant positive effect on cumulative abnormal returns.

In addition to earnings conference calls, a number of authors have also examined earnings press releases (Demers & Vega (2010, 2011), Henry (2006, 2008), Henry & Leone (2015), Davis et al. (2012), Davis & Tama-Sweet (2012)). However, given that this information tends to be disseminated via newswire texts, I have already discussed this literature in the previous section.

Several other studies have also investigated the transparency, or ‘readability’ of corporate reports. The main premise behind these studies is that reports that are more complex (typically measured by either document length, size of file, or some readability index) make it harder for investors to digest the information content contained therein. However, as these studies do not relate to textual sentiment, they will not be discussed here. The interested reader is referred to examples in Li (2010a) and Loughran & McDonald (2015).

Corporate disclosure material therefore tends to combine both quantitative and qualitative data, with the qualitative information providing an interpretation of the numerical data. Extracting this qualitative information can be challenging, and researchers typically focus on specific sections of corporate reports. The fact that reports tend to be split into well-defined sections, suggests that corporate disclosure sources are more structured than conventional news texts. Corporate disclosure material typically consists of the primary sources used by reporters to write news articles, and consequently the main advantage is that these materials can be used to gain an insight into

market fundamentals before the information hits the mainstream media. Furthermore, unlike conventional news sources, which can be thought of as ‘unexpected’ information, a large amount of corporate disclosure material (particularly earnings information and quarterly/annual reports) tends to be ‘scheduled’ or ‘expected’ information. Perhaps their most significant disadvantage is the frequency, and it can therefore be difficult to incorporate these information sources into traditional econometric methods. Nevertheless, these sources have proven useful in many studies.

2.5.3 Social Media

Another frequent source of information in the text analysis literature is social media texts. Social media texts have a much quicker publication cycle than conventional news texts, and the barrier to publication is also much lower. Indeed, anyone with a social media account can publish information, which is immediately received by anyone within their social network. As a result, information from social media accounts is of a much larger volume than other information sources. Studies involving social media texts therefore tend to utilise much larger corpora than studies that use other information sources. However, there are also several disadvantages to using social media as a source of investor sentiment. Because of the high volume of posts, studies which use social media texts tend to involve a much shorter time span than studies using conventional news. Financial markets rapidly evolve from one period to the next, and any results in one time period are unlikely to extend to another. Another disadvantage is that social media texts tend to be very noisy in the sense that they contain many abbreviations and acronyms, frequent misspellings, ungrammatical sentences (or indeed no sentence structure at all), or other non-Standard Internet vernacular. At best, these texts will simply increase the amount of pre-processing required. However, at worst these texts can prevent adequate parsing at all, and both dictionary-based or machine learning methods may struggle to cope with the non-Standard language. A further disadvantage with these text types is that because of the low barrier to publication and lack of regulation of these sources, there is a possibility that some of these texts are merely noise (e.g. rumours, exaggerations, misstatements) as opposed to carrying useful informational content that can help explain market variables. However, despite these criticisms a number of studies have found social media texts to be a useful source of investor sentiment.

Early studies of social media attempted to explore the effect of sentiment in financial message boards (Antweiler & Frank (2004), Das & Chen (2007), Kim & Kim (2014), Tumarkin & Whitelaw (2001). One of the first studies on financial message boards was Tumarkin & Whitelaw (2001), who examined the effect of sentiment on the website RagingBull.com against abnormal stock returns and volume. Using a corpus of 181,633 messages on stocks from the Internet service sector they found that message board activity neither predicted industry-adjusted returns nor abnormal trading volume, but that positive returns preceded days with strong positive opinions. Antweiler & Frank (2004) used a much larger corpus of 1.5 million messages about 45 companies in the DJIA index, posted on RagingBull.com and Yahoo! Finance message boards. The authors found that message posting helped predict volatility, and that an increasing level of disagreement amongst message posts induced higher volume trading. They also observed that an increase in the number of message board postings predicts negative next-day returns, but that this effect, while statistically significant, was economically small. Das & Chen (2007) also used messages from Yahoo! finance and, like Tumarkin & Whitelaw (2001), chose tech-sector stocks as his subject study. Using 145,00 messages for a two month period in 2001, he found that aggregate sentiment was positively related to aggregate stock index return on the following day. An example of a negative study is Kim & Kim (2014), who found no evidence that sentiment forecasts returns at either the aggregate or the firm-level. Using a corpus of 32 million messages from Yahoo! Finance message boards on 91 firms, they found that investor sentiment was in fact affected by prior stock price rather than vice versa. Whereas Tumarkin & Whitelaw (2001), Antweiler & Frank (2004), Das & Chen (2007) analysed the effect of sentiment at the aggregate level, rather than the firm-level, Kim & Kim (2014) analysed the effect at both the aggregate and individual firm-level.

Several studies have also explored the relationship between financial market activity and microblogs, including Twitter (Bollen et al. (2011), Zhang et al. (2012), Sprenger et al. (2014), Si et al. (2013a), Vu et al. (2012), Yu et al. (2013), Si et al. (2013b)) and finance-specific microblogging sites such as Stocktwits.com (Oh & Sheng (2011), Oliveira et al. (2013)). Perhaps the most widely-cited of these is Bollen et al. (2011), who found that changes in public mood (measured along 6 dimensions: Calm, Alert, Sure, Vital, Kind and Happy) was predictive of next-day returns of the Dow Jones Industrial Average. Si et al. (2013a) extended this work to obtain firm-specific tweets

for constituents in the S&P 100 index. The authors found a statistically significant relationship between extracted topic and future S&P 100 returns. Outside of Twitter, Oh & Sheng (2011) used a finance-specific microblogging website called Stocktwits, collecting 72,000 posts on almost 2,000 companies. The authors found that sentiment in microblogs had predictive power for both simple and market-adjusted returns. Microblog sites often have a limit on the length of the post²¹, rendering posts concise and to the point. Although this could lead to non-Standard vernacular (caused, in one sense, by users trying to limit their message size), it could also lead to posts that are highly focussed on a particular event. This of course contrasts with traditional news stories, which tend to be much longer and provide lots of non-sentiment background material.

A few studies have also been conducted using blogs. Gilbert & Karahalios (2010), for instance, used the blogging platform LiveJournal to extract emotional proxies of anxiety, worry and fear. The authors found that increases in anxiety led to downward pressure on the S&P 500 index. Yu et al. (2013) used a mixture of blogs, forums, news and microblogs. The authors here observed that positive sentiment in blog posts had a strong positive impact on returns whereas negative sentiment in forum posts had a negative impact on returns. Zhang & Skiena (2010) also exploit information from multiple sources, using a combination of newspapers, blogs and RSS feeds as their textual corpus. They found that sentiment in blog and microblog posts had less predictive power of future returns than sentiment expressed in news. This could be due to the fact that blogs and microblogs can be very noisy - perhaps in part due to their lack of regulation compared with conventional news media (Leinweber & Sisk (2011)).

2.6 Modelling & Statistical Estimation

The previous section discussed the main textual sources typically used in studies on sentiment analysis. While the system in this thesis is capable of using any unstructured textual information source, the case studies to be presented in Chapter 4 focus solely on news texts. Having chosen a suitable textual source, and constructed a sentiment proxy (typically using either a dictionary-based or machine learning approach), the

²¹At the time of writing, Twitter imposes a 140 character limit on each tweet's length.

next step is to estimate the effect of the sentiment proxy on a market phenomenon of interest. Usually this involves selection of appropriate market data, and an appropriate statistical method for examining the relationship between the proxy and the data. In this section I review some of the related work to show the breadth of market data and statistical methods that have been used in previous studies. This also provides background information required to understand the methodological approaches in the rest of this thesis. I also discuss some of the main published findings to date in order to show how previous works are related, and situate my own work within this existing body of knowledge. An overview of related studies and the various econometric methods used is given in Table 2.5.

Previous work in the financial sentiment analysis literature tends to focus on either aggregate market data (Tetlock (2007), García (2013), Bollen et al. (2011), Gilbert & Karahalios (2010), Wüthrich et al. (1998)) using an index such as DJIA or S&P 500, or else individual firm-level data (Ahmad et al. (2016), Tetlock et al. (2008), Loughran & McDonald (2015), Henry (2008)), such as company share prices or company earnings. Using aggregate market data is often advantageous, as it can be difficult to construct a continuous sentiment time series with news occurring every day. Few studies have explored the effect of sentiment at the sector-level (notable exceptions are Shynkevich et al. (2016) and Li et al. (2014)), and so this is an additional contribution that this thesis makes. Indeed, to the best of my knowledge this is the first study that has used a sector-level stock index, rather than a market-wide index such as DJIA or S&P 500, or constructing an index from an aggregate of individual share prices.

Regardless of using aggregate or firm-level data, the researcher also needs to determine the frequency of the market data's observations. A time series, by definition, is a sequence of equally spaced observations over time (Makridakis et al. (1998)). However the choice of spacing can vary enormously, and depends upon the market data being modelled. Most commonly, researchers work with daily market data (Tetlock (2007), García (2013), Ahmad et al. (2016), Ferguson et al. (2015), Das & Chen (2007), Ferguson et al. (2015), Bollen et al. (2011), Mao et al. (2015), Gilbert & Karahalios (2010)), as there is an abundance of historical daily data that is freely and easily available. In addition, daily data are usually the highest frequency at which news articles occur. More recently, researchers have become interested in high-frequency trading, where observations can occur at several times per day (Antweiler & Frank (2004),

Nassirtoussi et al. (2015), Schumaker & Chen (2006)). This type of *intraday* market data is usually reserved for textual sources that occur at a higher frequency, such as social media texts or message board posts. Market data can also occur over a longer period such as monthly (Chan (2003)), quarterly (Tetlock et al. (2008)) or annually (Li (2006), Loughran & McDonald (2015)). These studies usually attempt to forecast future earnings, often in response to earnings press releases or annual reports.

To model the data, there are a wide variety of econometric methods at the researcher's disposal. Approaches in the sentiment analysis literature are typically divided into either classification methods or regression methods. In classification methods (e.g. Bollen et al. (2011), Li et al. (2014), Oh & Sheng (2011), Nassirtoussi et al. (2015)), the aim is to forecast a change in the direction of a market variable, typically using either a binary classification structure ('rise'/'fall') or a ternary structure ('buy'/'hold'/'sell'). However, while these methods attempt to forecast a change in the direction of a given market variable, they do not attempt to forecast the magnitude of the change. Regression methods, on the other hand, attempt to forecast both the direction and the magnitude of the change and have been widely used in the econometric literature (Ahmad et al. (2016), Cook & Ahmad (2015), Davis et al. (2012), Tetlock (2007), García (2013), Groß-Klußmann & Hautsch (2011), Das & Chen (2007), Ferguson et al. (2015), Loughran & McDonald (2011), Henry (2008), Zhang et al. (2012), Antweiler & Frank (2004), Mao et al. (2015), Jegadeesh & Wu (2013)). This is a harder task than simple binary or ternary classification, but is naturally more suitable for the purpose of risk management. Consequently, regression methods form the basis of this thesis and so they are discussed in greater detail in the ensuing sections.

2.6.1 Linear Regression Models

One of the most widely used statistical models in econometrics studies is that of linear regression. Linear regression involves estimating the relationship between a (continuous) dependent variable y and one or more (continuous) explanatory x variables. The relationship between these variables is assumed to be linear in nature, and many existing studies (see Section 2.7) have observed a linear relationship between sentiment and various financial variables, such as changes in share prices and share volume. Recent research has begun to explore the non-linear relationship between sentiment and financial variables, but this is considered beyond the scope of this thesis. However, in the

regressions presented in Chapter 4, each of the continuous variables is log transformed to help ensure that the variance is constant. A by-product of this transformation is that it removes some of the non-linearity between the variables under investigation.

Thus the linear relationship between the variable y and one or more x variables can be represented via the following equation:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon_i, \quad i = 1 \dots n \quad (2.2)$$

Where:

β_0 = a constant term denoting the expected value of y when all x variables are set to 0.

$x_1 \dots x_p$ = the p independent variables

$\beta_1 \dots \beta_p$ = the p model coefficients (weights of the x variables)

ε = a random error term (assumed to have a mean of 0)

Each independent variable $x_1 \dots x_p$ may have a different level of importance in determining the value of the dependent variable y . The model coefficients $\beta_1 \dots \beta_p$ determine the level of importance of the associated x variable. A coefficient close to zero suggests that the associated x variable has little weight in determining the value of y , whereas a higher coefficient suggests that the associated x variable is more useful for describing changes in y .

There are a variety of methods that can be used to estimate the regression coefficients, however one of the most widely used methods is the ordinary least squares (OLS) algorithm²². The OLS algorithm estimates the model's 'best' coefficients by minimizing the squared difference between the actual and fitted values:

$$OLS = \text{minimize} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.3)$$

²²Other examples include least absolute deviations, maximum-likelihood estimation, and least trimmed squares.

Where:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$ (the estimated value of y_i)

y_i = the actual observed value of y_i in the dataset

The objective of the OLS procedure is therefore to minimize the residual error term ε for each observation i . After the model's coefficients have been estimated, the fitted linear equation can subsequently be used to extrapolate *unknown* values of y given known values of the x variables. Alternatively, the values of the coefficients can be examined to explore the strength of the relationship between the y variable and each x variable, to see which x variable contains any useful information about the value of y .

2.6.2 Autoregression

In time series data it is often the case that the value of a variable depends upon its own historical values. This dependency can be encapsulated formally within the regression model posited in Equation 2.2. Rather than the variable y being dependent upon exogenous explanatory x variables, it is represented as a function of past values of itself. Models of this form are known as *autoregression* models. Formally, an autoregression model of order p (denoted as $AR(p)$) can be expressed as:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1 \dots n \quad (2.4)$$

Where:

β_0 = a constant term

$\beta_1 \dots \beta_p$ = the p th autoregressive coefficient

$y_{t-1} \dots y_{t-p}$ = the lagged values of y at time $t - p$

t = the current time period

ε = a random error term at time t (assumed to have a mean of 0)

Alternatively, in short-hand notation this can be expressed as:

$$y_t = \beta_0 + L_p y_t + \varepsilon_t \quad (2.5)$$

Where:

$$L_p y_t = \sum_{i=1}^p \beta_i y_{t-i}$$

In this shorthand notation the lag operator L_p converts a variable y_t into a row vector of its lagged (historic) values $[y_{t-1}, \dots, y_{t-p}]$. Thus in an autoregression model the dependent variable is a function of its lagged values, plus a stochastic error term.

The model parameters can be calculated in the same manner as the ordinary regression model. Once the parameters of the model have been estimated (and assuming that they will remain stable), the model can then be used for forecasting. If t denotes a time period in the future for which data is not yet available, then an estimate for y_t can be obtained by substituting the historical values of y for the lags $y_{t-1} \dots y_{t-p}$.

Like the ordinary regression model in Equation 2.2 the scope of an autoregression model is not limited to forecasting. The autoregressive coefficients can also be used to examine the strength of the relationship between the past and current values of y : specifically, what information do the known lagged values of y have about its unknown future value.

2.6.3 Vector Autoregression

The vector autoregression (VAR) model is a natural extension to the autoregression previously described. The VAR model, first popularised by Sims (1980), suggests that a variable can be dependent not only its past history, but also on the history of other (exogenous) variables. Unlike the ordinary autoregression model, in which a single variable y_t is modelled in terms of its past values, in a vector autoregression model several variables are modelled in terms of their own past and the past of other variables. The term ‘autoregression’ refers to the fact that lagged values of the dependent variable are used as the explanatory variable, and the term ‘vector’ refers to the fact that the variables being modelled are typically represented in vector form.

To explain the core principle behind the VAR model let us consider the simplest case - a bivariate VAR involving only two variables $y_{1,t}$ and $y_{2,t}$. The current value of each of these variables at time t depends upon the historical (lagged) values at time $t - 1 \dots t - p$. While any number of lags can be chosen, for simplicity here only one lag of each variable is shown:

$$\begin{aligned}
y_{1,t} &= \beta_{1,0} + \beta_{1,1}y_{1,t-1} + \beta_{1,2}y_{2,t-1} + \varepsilon_{1,t} \\
y_{2,t} &= \beta_{2,0} + \beta_{2,1}y_{2,t-1} + \beta_{2,2}y_{1,t-1} + \varepsilon_{2,t}
\end{aligned}
\tag{2.6}$$

Where:

$\varepsilon_{1,t}/\varepsilon_{2,t}$ = an error term at time t with $E(\varepsilon_{1,t}) = 0, E(\varepsilon_{2,t}) = 0, E(\varepsilon_{1,t}, \varepsilon_{2,t}) = 0$

Thus each variable in the model has its own equation. The above VAR model can be more compactly represented using the following matrix notation:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \beta_{1,0} \\ \beta_{2,0} \end{bmatrix} + \begin{bmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}
\tag{2.7}$$

Where the model's variables, error term and constant term are represented in vectors and the coefficients are defined in matrices.

The above model was an example of a VAR(1) model with 2 variables²³. In reality a VAR model would usually contain a larger number of variables and indeed is not confined to only one lag of each variable. The general form of a VAR model with k variables and p lags can be expressed in matrix notation as follows:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} \beta_{1,0} \\ \beta_{2,0} \\ \vdots \\ \beta_{k,0} \end{bmatrix} + \begin{bmatrix} \beta_{1,1}^1 & \beta_{1,2}^1 & \cdots & \beta_{1,k}^1 \\ \beta_{2,1}^1 & \beta_{2,2}^1 & \cdots & \beta_{2,k}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1}^1 & \beta_{k,2}^1 & \cdots & \beta_{k,k}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \dots + \begin{bmatrix} \beta_{1,1}^p & \beta_{1,2}^p & \cdots & \beta_{1,k}^p \\ \beta_{2,1}^p & \beta_{2,2}^p & \cdots & \beta_{2,k}^p \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1}^p & \beta_{k,2}^p & \cdots & \beta_{k,k}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{k,t} \end{bmatrix}
\tag{2.8}$$

Where each of the model's k variables has its own regression equation. The coefficients of the regression can be computed using the OLS procedure described in the

²³The notation VAR(p) denotes a vector autoregression model with p lags (historical values of each variable).

previous section. While the number of lagged variables could vary for each of the variables, in practice the same number of lagged terms are used for each equation.

One of the primary advantages of the VAR model is that it allows for the detection of *Granger causality*. Granger causality (Granger (1969)) implies that if a variable x ‘Granger-causes’ another variable y , then lagged values of x contain information about y above and beyond any information contained in the past values of y alone. Thus computing a series of regression models (one for each variable) allows one to observe the nature and extent of this Granger-causality effect and, importantly, whether the effect is *bidirectional*.

The vector autoregression model has been used in a multitude of studies, including Mao et al. (2015), Tetlock (2007), Tetlock et al. (2008), García (2013), Ahmad et al. (2016), Cook & Ahmad (2015), Groß-Klußmann & Hautsch (2011), Brown & Cliff (2004), Ferguson et al. (2015), Si et al. (2013a), Tumarkin & Whitelaw (2001). These studies attempt to forecast a particular market variable based upon a vector of lagged (historic) variables. These lagged variables typically comprises historic values for the dependent variable, the sentiment variable, and other market variables known to impact the dependent variable. Thus the market variable is posited as a linear function of its historic values, and the historic values of other variables (Makridakis et al. (1998)). The VAR model is used extensively throughout this thesis, and therefore will be described in greater depth in Chapter 3.

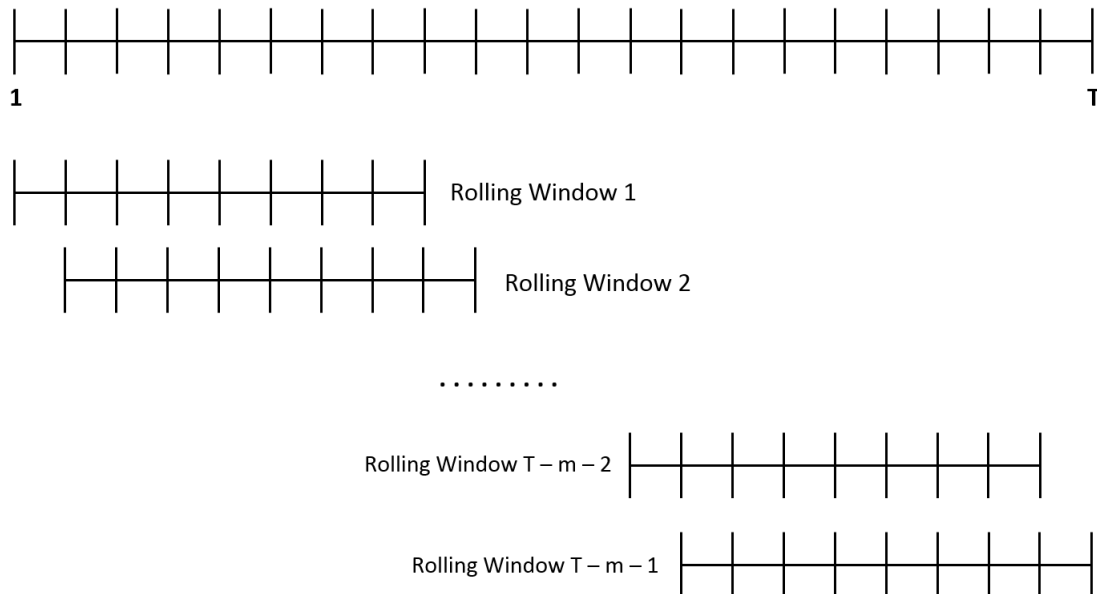
2.6.4 Rolling Regression

Extrapolating future (out-of-sample) values assumes that the model’s parameters remain stable over time. However, financial markets are dynamic in nature and any statistical model that attempts to capture their variability will inevitably be out-of-date as the market environment changes. For instance, García (2013) argues that the relationship between sentiment and index-level returns is stronger during recessions than during expansionary periods. The adaptive market hypothesis (Lo (2004)) proposes that financial markets evolve to reflect changing environmental conditions and changing participants. The consequence of this dynamic marketplace are that regression coefficients calculated over a given time period may not hold for future time periods.

To explore the stability of the model’s parameters over time, or indeed model the dynamic nature of the marketplace more generally, one could partition the sample

period into a sequence of sub-samples, computing a new model for each period. Stability of model parameters will be reflected in similar estimates across successive time periods. A common approach in time series analysis is to use a rolling methodology, whereby the sub-samples are taken sequentially over time. The main specifications for a rolling method are the window size, representing the size of the sub-sample, and the increment size, denoting how many time periods the rolling window should be shifted forward. Figure 2.1 shows the general schematic of the rolling methodology.

Figure 2.1: Intuition behind the rolling window. The figure here is adopted from MathWorks (2014). The data is split into $T - m - 1$ sub-samples, where T is the total number of observations in the time series and m is the width of the rolling window. The window here is shifted (rolled) forward one time period at a time, although it could be shifted forward by any constant number of periods. During each rolling window a particular function is computed using the sample data within the current window. Note that during each rolling window, one of the observations from the previous window is lost, and another is gained.



While any econometric method could be computed during each rolling window, this thesis is concerned with computing a rolling VAR model. During each rolling window a new VAR model is computed using market and sentiment data from that particular sample. The window is then rolled forward by one day and a new VAR model is computed in the new period. The model's coefficients during each time window capture the (in)stability of the model over time: if the coefficients are similar for all time periods then the model's estimates are considered to be constant (i.e. time-invariant). However, if the market environment changes this could cause the relationship between market (or sentiment) variables to change also, reflecting a change in the estimated model

coefficients.

The key parameter in the rolling model is the window size, and different window sizes will yield different model estimates. A wider window size will take longer to adjust to changing market dynamics, and extreme observations (outliers) will tend to persist. In addition, a wider window results in an averaging effect whereby potentially illuminating effects during a particular time period will be concealed. In contrast, a shorter window size will quickly adjust to reflect new market conditions, and extreme observations will quickly exit the rolling window. Using a shorter window also reduces the number of observations per sample, reducing the power of any statistical tests and increasing the variability in the dataset. Ultimately, the window size depends on the nature of the problem, and the literature offers no clear suggestions as to the appropriate solution. Existing work has used various prediction windows, ranging from two weeks in Jin et al. (2013) to a full trading year in Ahmad et al. (2016).

One of the limitations with the rolling regression methodology is the issue of multiple comparisons. While a single regression test carries a relatively small probability of a Type 1 (false positive) statistical error, conducting multiple hypothesis tests simultaneously (here one per rolling window) substantially increases this risk. Thus as the number of simultaneous hypothesis tests increases, so too does the probability of encountering a significant result purely by chance. One way of controlling for this *family-wise error rate*, is to use a corrective procedure such as Bonferroni correction. Bonferroni correction works by adjusting the significance threshold (α) downwards, typically by dividing α by the number of hypothesis tests to be conducted. This results in a smaller critical value, requiring stronger evidence to reject the null hypothesis. However, by decreasing Type 1 errors this invariably leads to an increase in the number of Type 2 errors - failing to identify a significant effect when one is present. Because rolling regression involves the computation of hundreds (or potentially thousands) of statistical tests, this results in an extremely low significance threshold. For instance, if one adopts the standard significance level of $\alpha = 0.05$, and computes 1000 hypothesis tests, the Bonferroni-corrected significance level would be $\alpha = 0.00005$ - a very small number indeed. This means that one would have to obtain an extremely low p-value in order to reject the null hypothesis. It could also be argued that in the context of time series models, Bonferroni correction could be considered overly conservative (Blackford et al. (2009)). In the rolling regression models used in this thesis, each rolling window

overlaps with the preceding and following windows introducing a level of dependency between tests. Conversely, Bonferroni adopts a worst-case scenario that assumes complete independence between hypotheses tests. In this thesis I therefore do not make any family-wise error corrections to account for the multiple comparisons problem introduced by the rolling regression models. Although this approach appears to mimic that adopted in previous studies employing rolling regression models (e.g. Ahmad et al. (2016)), it is identified as a limitation of this research. Future work should therefore attempt to address this issue further, by making efforts to reduce the family-wise error rate, and strengthen the statistical power of the results.

2.6.5 Panel Regression & Event Studies

Studies on firm-level sentiment analysis pose a particular challenge, as news does not occur every day even for the largest of firms. Consequently, it can be difficult to construct a continuous sentiment time series without gaps from missing observations. However, there are a number of ways to circumvent this issue. Firstly, one could simply pool all the firm-level observations together into a single time series and use pooled regression as in Tetlock et al. (2008) and Das & Chen (2007). However, while this approach requires estimation of few additional parameters, it assumes that the variables in the model are constant between firms (Brooks (2014)). This latter assumption is highly unlikely, and it has been observed in the literature, for instance, that sentiment impacts small firms differently from larger firms (Baker & Wurgler (2006)). A second approach, which does allow for variation between firms, is to use a panel-based regression model as in Ahmad et al. (2016), Zhang et al. (2012) and Antweiler & Frank (2004). Typically, panel-based studies use a firm fixed-effects model, whereby the constant in the regression model is allowed to vary cross-sectionally (i.e. between firm), but the other market variables are fixed both cross-sectionally and over time. This is typically accomplished by including a dummy variable for each firm in the study. If the firms are large enough, and an appropriate data source is selected, then it may be possible to run separate VAR models for each of the firms. However, using firm-level VAR models would not have the same statistical power as using pooled or panel-based models: in the latter cases the observations from each individual time series would be combined into a large, single time series, increasing the number of observations available for statistical analysis.

Another regression model often used for firm-level analysis is the event study method-

ology (see for instance Henry (2008), Jegadeesh & Wu (2013), Feldman et al. (2010), Davis et al. (2012)). In event studies, one attempts to predict the (cumulative) abnormal returns around a particular event, such as a corporate merger or the filing of a corporate report. Specifically, one investigates whether a stock exhibited higher (lower) returns around an event than would have been expected in the absence of the event. The abnormal returns are usually calculated as the difference between the stock's return and a given stock index or collection of stocks.

2.7 Key Market Variables

To date a wide variety of economic variables have been modelled in the literature. This section reviews some of the work exploring the relationship between sentiment and these variables.

Many related studies have attempted to model the relationship between changes in sentiment and changes in returns (Tetlock (2007), Ferguson et al. (2015), Henry (2008), García (2013), Ahmad et al. (2016), Bollen et al. (2011), Li (2006), Tetlock et al. (2008), Mao et al. (2015), Jegadeesh & Wu (2013), Gilbert & Karahalios (2010), Chen et al. (2014), Das & Chen (2007), Antweiler & Frank (2004)), where returns denote the change in price between two consecutive periods. The majority of these studies have found that sentiment forecasts returns of various time frequencies²⁴, often with subsequent reversal (Tetlock (2007), Ahmad et al. (2016), Mao et al. (2015), Ferguson et al. (2015)) suggesting a transient effect. Most researchers have focused purely on the effect of negative sentiment (Ahmad et al. (2016), Tetlock (2007), Loughran & McDonald (2011), Zhang et al. (2012), Chen et al. (2014)), though some researchers have looked at positive sentiment in isolation (Henry (2008), Oh & Sheng (2011)), both negative and positive sentiment together (García (2013), Jegadeesh & Wu (2013), García (2013)) or a net, aggregated measure usually expressed as the (normalized) difference between the two measures (Li et al. (2014)). In García (2013) and Mao et al. (2015) it was observed that the effect of sentiment is stronger during recessionary and turbulent market periods, and in Ahmad et al. (2016) it was observed that the role of sentiment is episodic. This time-varying impact of sentiment is interesting, as it would

²⁴Notable exceptions are Das & Chen (2007) and Antweiler & Frank (2004), who both observe no significant relationship between returns and sentiment expressed in message boards postings.

seem to suggest that while markets can be efficient there are certain periods where when they are inefficient, supporting both the efficient market theory and behavioural finance paradigms.

A number of researchers have also explored the relationship between sentiment and a firm or index's volume. The volume is often used a proxy for market liquidity, and is typically represented by the number of shares exchanged during a given time period. Tetlock (2007), Loughran & McDonald (2011), Zhang et al. (2012) all find that negative sentiment has a significant impact on volume, suggesting that an increase in negative sentiment leads to a higher number of trades.

Some studies have also focussed on forecasting a firm or index's volatility - the amount of variation in share prices over time. A simple measure of volatility is the standard deviation, which measures how far a given share prices deviates from its mean value. Volatility is often viewed as a measure of financial risk, as periods of high volatility mean that prices are rapidly rising or falling during this time, and are thus less predictable than during periods of low volatility. In the sentiment analysis literature, Loughran & McDonald (2011) and Zhang et al. (2012) have observed that negative sentiment directly impacts firm-level volatility, implying that an increase in negative sentiment leads to an increase in volatility.

Other market variables of interest include forecasting quarterly earnings (Tetlock (2007)), annual earnings (Li (2006)) and currency pair movements (Nassirtoussi et al. (2015)), however a discussion of these variables is deferred here as they are not directly relevant to this thesis. In addition, some studies have also examined the effect of *news flow* - the number of texts during a given period - on a number of market variables (e.g. Groß-Klußmann & Hautsch (2011) and Antweiler & Frank (2004)). However, given that this study focuses on the effect of sentiment rather than news flow, a discussion of these studies is beyond the scope of this thesis.

A summary of some of the existing literature's key findings is shown in Table 2.5. This table lists a variety of the different market data that have been studied at various time frequencies. Additionally, this table shows some of the core statistical methods that have been used, summarising the studies described in the previous section.

Table 2.5: Summary of pertinent literature grouped by frequency of market data.

Ref.	Market Data	Frequency of Market Data	Method	Key Findings
Klußmann & Hautsch (2011)	39 LSE Stocks	Intraday	Regression (VAR)	Sentiment predicts future price trends; arrival of news influences volatility, money value traded, average trade size and bid-ask spreads
Antweiler & Frank (2004)	45 DJIA stocks	Intraday	Regression (Panel) & GARCH	Higher number of messages predicts negative subsequent returns; bullishness affects neither returns nor volatility
Schumaker & Chen (2006)	S&P 500 stocks	Intraday	Support Vector Regression	Compared models using bag-of-words, named entities and noun phrases; found certain noun phrases could predict financial returns
Nassirtoussi et al. (2015)	FOREX	Intraday	Classification (SVM, KNN, Naïve Bayes)	A predictive relationship exists between news headlines and currency-pair-price movements
Ahmad et al. (2016)	20 stock returns	Daily	Regression (Panel & Rolling Window VAR)	Firm-level negative sentiment episodically impacts firm-level returns; the effect is sometimes reversed
Tetlock (2007)	DJIA Index	Daily	Regression (VAR)	Negative sentiment influences next-day returns with subsequent reversal; sentiment also impacts trading volume
García (2013)	DJIA Index	Daily	Regression (VAR)	The effect of negative sentiment on next-day returns is stronger during recessions and on days after weekends or public holidays
Das & Chen (2007)	24 Tech Stocks / MSH Index	Daily	Regression (Pooled)	MSH index weakly related to prior day's sentiment index; individual firm-level returns not affected by prior day's sentiment index
Ferguson et al. (2015)	FTSE 100 Firms	Daily	Regression (VAR)	Tone (positive and negative sentiment) and newsflow impacts future stock returns; the impact of volume is more pronounced than tone
Zhang et al. (2016)	100 S&P stocks	Daily	Regression (Panel)	Increased negative sentiment influenced volatility and volume, but not returns; different sentiment dictionaries give conflicting results

Ref.	Market Data	Frequency of Market Data	Method	Key Findings
Mao et al. (2015)	4 major stock indices	Daily	Regression (VAR)	Twitter bullishness predicts stock index returns with subsequent reversal. Effect of bullishness is stronger in extreme market conditions
Chen et al. (2014)	7,000 firms	Daily	Regression	Negative terms in firm-level news impact future abnormal returns
Gilbert et. al. (2010)	S&P 500 Index	Daily	Granger Causality	Increase in anxiety index leads to lower subsequent returns
Oh & Sheng (2011)	72,211 stocks	Daily	Classification (J48)	Bullish sentiment in microblogs predicts market-adjusted returns
Bollen et al. (2011)	DJIA Index	Daily	Classification (Fuzzy Neural Network)	Changes in public mood on Twitter predict directional changes in DJIA
Li et al. (2014)	22 Hang Seng stocks	Daily	Classification (SVM)	Find GI & LM dictionaries outperform bag-of-words model in prediction of firm-level, sector-level and industry-level returns.
Davis et al. (2012)	23,017 firm quarterly returns	3-day window	Event study	Optimism in earnings press releases positively impacts abnormal returns around issue of press release; future firm performance is also affected
Henry (2008)	562 stocks	3-day window	Event Study	Positive sentiment leads to higher returns around earnings press releases
Feldman et al. (2010)	8,219 firms	3-day window	Event Study	Change in tone in 10-K and 10-Q reports leads to higher abnormal returns
Jegadeesh & Wu (2013)	7,616 firms	3-day window	Event Study	Weighted sentiment measure (incorporating both positive and negative terms) impacts abnormal returns around Form 10-K filing days
Tetlock et al. (2008)	Earnings info.	Quarterly	Regression	Negative sentiment forecasts low quarterly earnings for 1,063 firms
Li (2006)	Earnings Reports	Annual		The words ‘risk’ and ‘uncertain’ predict low annual earnings
Loughran & McDonald (2011, 2015)	8,341 firms	Annual	Regression (OLS, Logit)	Negative sentiment in annual reports impacts trading volume, returns, unexpected earnings, and volatility

2.8 Summary

This chapter began by motivating the role of investor sentiment in the context of determining asset prices. The efficient market hypothesis was presented as one of the dominant asset pricing theories over the past 50 years. According to this hypothesis, the price of an asset reflects all known information about that asset. As a result it not possible for an investor or trader to earn an above-average profit except by chance. However, the efficient market hypothesis assumes that investors have equal access to the same information and that they are equally skilled in the processing of this information. This is an unrealistic assumption, as markets are full of a wide range of participants, each with access to different information sets, and each with their own prior beliefs and biases. The efficient market hypothesis has also been criticised on its inability to explain a number of market anomalies. As a result, the field of behavioural finance has been presented as an alternative paradigm to efficient market theory. Behavioural finance suggests that financial markets are not always efficient, and that at certain times investors' decision making is driven by behavioural or psychological factors. This can lead to irrational decision making whereby investors overvalue or undervalue the price of an asset, ignoring information about its fundamental value.

The literature on sentiment analysis claims that an insight into investors' thought processes can be gleamed by analysing investor sentiment. Naturally, it is not possible to ask every market participant what they are thinking and, even if it were, this would be likely to change in an instant. Instead, existing research on sentiment analysis has turned to indirect measures of investor sentiment in the form of *sentiment proxies*. Rather than conducting expensive and timely sentiment surveys, it was argued that a sentiment proxy could be constructed by analysing the texts typically read by investors. Principally, this is accomplished by analysing the positive and negative affect in text, which functions as a proxy for bullish and bearish investor sentiment. Thus the task of sentiment analysis amounts to deriving a method for quantifying this positive or negative affect.

Typically the methods for quantifying affect are divided into two main approaches: dictionary-based methods and machine learning methods. In dictionary-based methods a pre-compiled dictionary of sentiment-bearing lexical items is used to count the number of words and phrases. Contrarily, machine learning methods attempt to induce a set

of features from the data that are indicative of sentiment. Both approaches have their own advantages and disadvantages, which were discussed in Section 2.3.3.

Having discussed various methods for computing sentiment, the discussion then progressed to relevant literature in the area of sentiment analysis and financial markets. A discussion of the textual sources typically used in sentiment analysis ensued, with pertinent literature from each source. Following this, attention turned to a discussion of the main econometric and statistical methods used in studies of sentiment analysis. Here, the various types of market data were discussed, along with the frequency of data and the typical methods used. Particular attention was paid to the topic of regression analysis, as this forms a fundamental role in this thesis. Finally, the chapter ended with a discussion of the key findings to date, giving an overview of the various market data that has been used.

Throughout this chapter, and indeed this thesis more generally, it has been suggested that the morally vicious and virtuous terms are statistically associated with the negative and positive terms. In this chapter I presented initial evidence of this, documenting how the vice and virtue terms in the General Inquirer dictionary (one of the major dictionaries used in studies on sentiment analysis) comprise a subset of the negative and positive terms respectively. Given this observation, one wonders whether this ethical subset of evaluative terms can be used as a proxy for investor sentiment. To this end, I develop a sentiment analysis system which is capable of automatically extracting evaluative and ethical sentiment proxies from text, and using these proxies in an econometric model to explore their impact on financial returns. In the next Chapter I provide a detailed overview of this system.

Chapter 3

Methods

3.1 Introduction

News and other information sources are integral to financial markets. Textual data can provide factual information that supports, contradicts, or is otherwise unavailable in quantitative market data. However, textual data can also reveal the general sentiment of the market. This so-called ‘market sentiment’ reveals investors’ opinions and attitudes towards a particular security, industry or indeed the market as a whole. By perusing these texts an insight into the investor thought process can be gleaned. This insight can subsequently be exploited to determine whether an asset is undervalued or overvalued relative to its fundamental value. However, with the barrage of information facing the typical investor, automated systems are needed to extract and analyse this information from unstructured texts. This chapter describes the development of one such system.

In this chapter a system is presented for automatically extracting and analysing investor sentiment from unstructured text. The purpose of this system is to explore the relationships between affective language in textual data and stock price movements. However, practically the system could also be used as a decision support system for investors to assist with the processing of information. In particular, the system aims to simulate the tasks of data collection, perusal, aggregation and inference that are typically conducted by investors when reading the news. However, unlike human decision making, which is subject to emotional and psychological biases, the system presented here offers an automatic and consistent method that is significantly faster at processing information. Furthermore, the system is scalable, in that it can process a multitude of

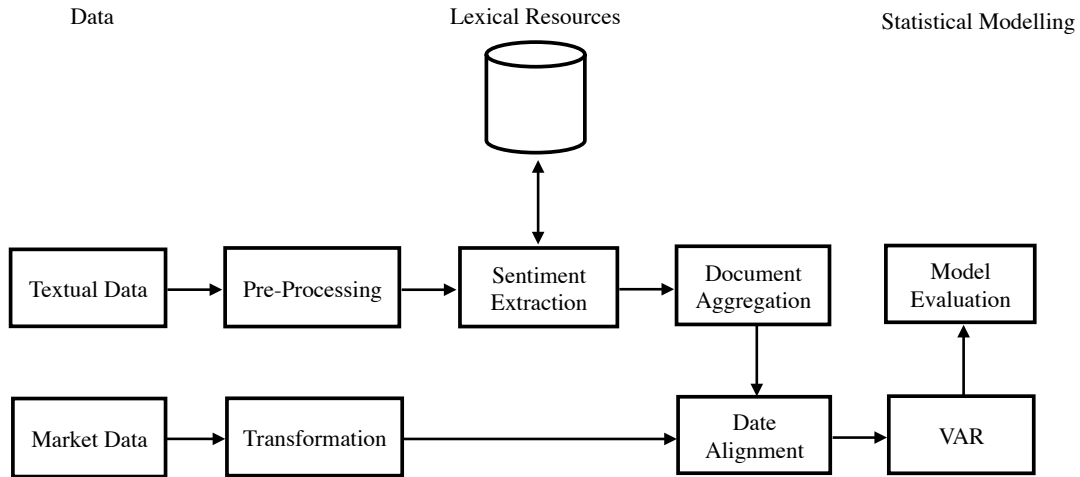


Figure 3.1: Key components of the sentiment analysis system.

documents without succumbing to physical obstacles such as fatigue.

The system's architecture can be broadly divided into two primary components:

1. A text analysis component (written in Java)
2. A statistical modelling component (written in R)

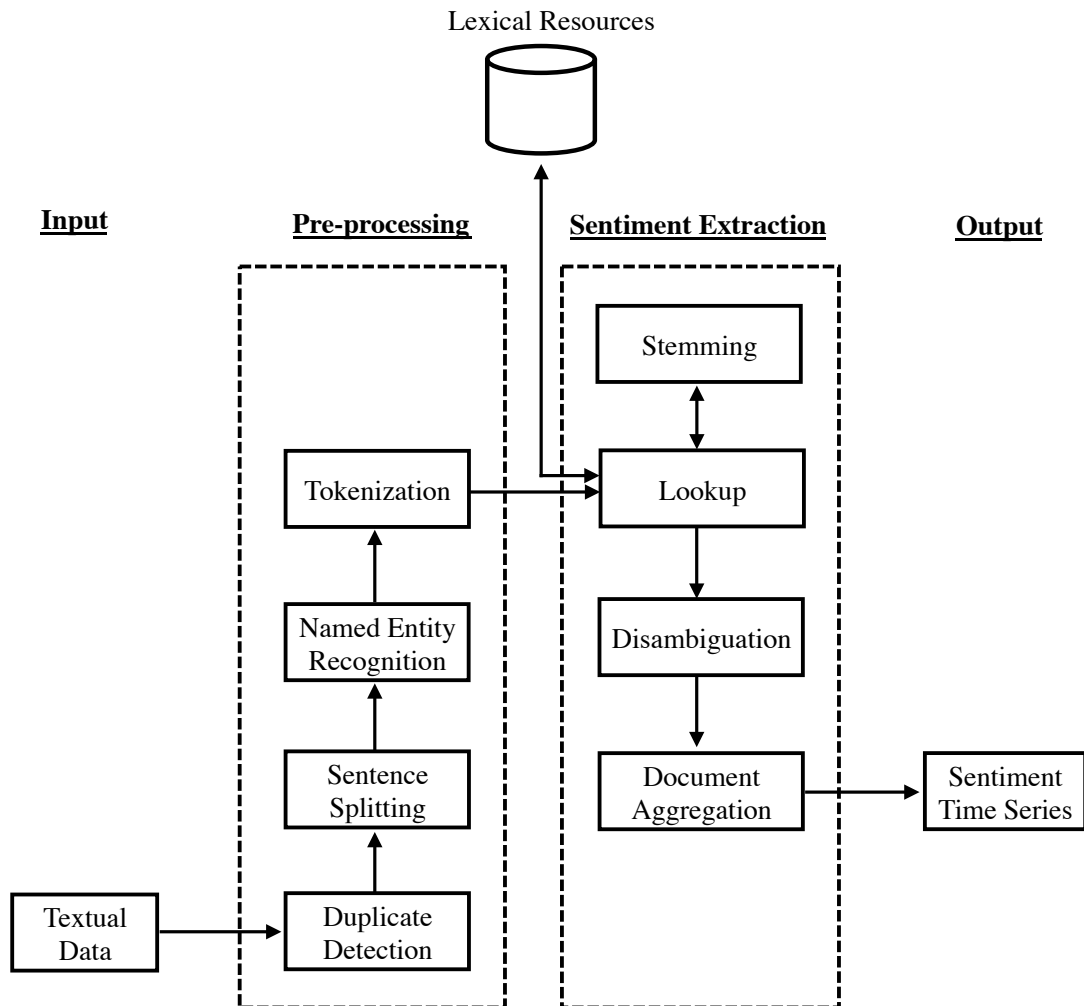
The text analysis component takes a corpus of unstructured textual data as input, pre-processes and parses this input, and produces a quantitative sentiment time series as output. The statistical modelling component takes the sentiment time series and a set of financial returns as input, and produces a statistical model describing the interrelationships between the sentiment proxy and financial returns. An overview of the system's architecture (previously mentioned in Chapter 1) is shown in Figure 3.1.

The main benefit of the system is its ability to handle a wide variety of textual and financial data. While news texts form the basis of the case studies presented in Chapter 4, in reality the system is capable of producing a statistical model from any unstructured textual data annotated with temporal information. Likewise, while returns form the basis of this study, with minor refinements the system is able to analyse the relationship between sentiment and other market variables, such as volume or volatility. In the rest of this chapter I describe each of the system's components from Figure 3.1 in more detail.

3.2 Textual Analysis

In this section I describe the individual steps that make-up the text analysis pipeline. In particular, I describe how the system takes qualitative textual data as input, pre-processes and parses this input, and converts it into a quantitative sentiment time series that can be used for statistical modelling. An overview of this text analysis component is given in Figure 3.2.

Figure 3.2: Key components of the textual analysis procedure designed in this thesis.



The system’s textual analysis pipeline essentially consists of two components:

- A pre-processing component in which the input textual data is cleaned, transformed, and tokenized into a form that can be used in the next stage;
- A sentiment extraction stage, which extracts affect or ethical information from the pre-processed textual data and outputs a time series;

The key component of the textual analysis system is the sentiment extraction stage. During this stage a lookup is performed in the GI dictionary for each of the words in a given text. If a word is found in the dictionary, then it is annotated with the entry’s associated tags ²⁵.

The remainder of this section describes each of the sub-components in greater detail. While each of these subcomponents will be described independently, the reader should note that the components are largely interwoven. Within the pre-processing stage, the tokenization, named entity and sentence splitting tasks are all completed concurrently using the Stanford Core NLP toolkit (Manning et al. (2014)). Likewise, the lookup, stemming and disambiguation tasks are also tightly interconnected. Aside from the obvious advantage of increasing processing efficiency, this also reduces the amount of failure points within the system.

3.2.1 Textual Data

One of the most important issues in text analysis is the selection of good quality data. One often hears the dictum ‘*garbage in, garbage out*’, reflecting the notion that poor-quality input will yield poor-quality output. This is especially true when working with linguistic data, which is rife with ambiguity, inconsistencies and inaccuracies. Attempting to derive intelligent information from this noisy, unstructured text is a challenging task by itself, and is exacerbated by poor quality input. Consequently, this section describes the textual data that is taken as input by the system, and discusses some of the ways to ensure its quality.

The textual data expected by the system is of the form of unstructured texts paired with temporal information. Specifically, the system assumes that the filename of each

²⁵Note that while the system annotates a word with all of its associated GI tags, the focus of this thesis is on the affect and ethical categories previously described.

input text is of the format *dd.mm.yyyy@*.txt* where ‘dd’ denotes the day of the month, ‘mm’ denotes the month, ‘yyyy’ denotes the year and * denotes one or more wild-card characters used to distinguish filenames of the same date. The dates themselves can be of any sequential time period (e.g. daily, weekly, monthly or yearly), however this thesis focusses solely on the use of daily texts.

The input texts can also originate from any source (e.g. news, blogs, company reports), with the primary condition that they are written in Standard English²⁶. Existing research has typically analysed news and other textual sources at one of three levels of market granularity:

- Market-level: news about a particular market, with a mixture of companies and sectors.
- Industry-level: news about a particular market industry or sector, such as news about commodities.
- Firm-level: news about specific companies.

The system here is capable of dealing with news at any of these three types of market granularity, however the main focus of this thesis is on firm-level news. This presents two major challenges that are not present with more general types of news. The first challenge is how to ensure that a given news article is written about a firm, as opposed to merely mentioning the firm in passing. The second challenge is providing adequate coverage of a particular firm, as even the largest firms do not necessarily have news written about them on every trading day.

The first challenge - ensuring that an article is written specifically about a particular firm - is an area of research known as document *aboutness*, and poses an interesting research topic by itself. Existing research on sentiment analysis (e.g. Ferguson et al. (2015), Shynkevich et al. (2016), Sinha (2016), Lavrenko et al. (2000)) typically uses external relevance assignments - documents which have been hand-tagged with the company’s name, usually by either the company themselves or by the data provider that administers access to it. The obvious limitation of this approach is that it makes

²⁶This requirement comes from the fact that the lexical resources used by the system are written in standard American English. Functionality also exists within these resources to cope with British English spellings. A major performance degradation could occur from using non-standard English sources, such as social media or other sources with frequent misspellings.

replication of existing studies very difficult, as the same labelled data will be needed in order to replicate any experiments. Extending existing studies will also be very difficult as there is no guarantee that the labelling conventions remain constant over time²⁷. To prevent a system being limited to one particular type of news source, an alternative method of ensuring document aboutness is needed.

A number of authors have defined their own selection criteria to determine document aboutness. In Tetlock et al. (2008), for instance, the authors require the firm's name to appear at least once within the first 25 words (including the headline), and at least twice in the body of the story. Ahmad et al. (2016) require each article to contain the company's name at least once in the headline, with at least five additional mentions in the body of the text. Intuitively, it seems that the firm name occurring in the headline and repeated mentions in the text are strong indicators of document aboutness. This indeed is confirmed in the aboutness literature, which tend to use these features in machine learning algorithms to classify document aboutness (e.g. Paranjpe (2009), Gamon et al. (2013)).

In this thesis the approach in Tetlock et al. (2008) was found to be too imprecise, as several of the filtered documents in the firm-level and industry-level case studies were not really written about the firms being researched. Conversely, the criteria in Ahmad et al. (2016) was found to be *too* restrictive, filtering out many very relevant documents that were in fact about the companies being studied. As a result, for the purpose of this thesis I decided to adopt a mid-point between the two studies. Specifically, I require that each document contains the firm's name (or common abbreviations) in the title of the article, along with at least 3 additional mentions in the body of the text. While this might exclude some articles that are potentially relevant, it maximizes the chance that the remaining articles are written about the company of interest, and was found to be less restrictive than the approach adopted in Ahmad et al. (2016). The obvious advantage to this approach is that the system is capable of working with any unstructured text, without being dependent upon metadata that is tailored to a specific type of document. However, note that this approach of ensuring document aboutness does assume that each document has an associated title. While this assumption is

²⁷ A personal conversation with an employee from ProQuest (a major aggregation service of historical news items) revealed that labelling conventions change fairly frequently, and are not consistent either between sources, or within the same source.

likely to hold for most formal news texts, it may not hold for other text types such as social media texts. The area of document aboutness would be an interesting area for future work, and additional procedures to filter out irrelevant documents could easily be added to the system developed in this thesis.

The specified selection criteria enables the use of any textual data that is annotated with temporal information. In this thesis all textual data was downloaded manually from the ProQuest or LexisNexis news archives²⁸. Although it would be possible to automate this process, such a task would require a perpetual licence subscription in order to access the requisite APIs. Both the ProQuest and LexisNexis archives offer sophisticated search functionality that enable the user to specify precise source information, temporal information, combinations of keywords, and to search in various subsections of the documents²⁹. Both services also place restrictions on the format of the articles that can be exported. In this case I exported documents from LexisNexis in a textual format, and documents from ProQuest in an HTML format as these were deemed the easiest formats to parse.

A bespoke LexisNexis and ProQuest parser (created in Java) was written by the author to parse each of the ProQuest and LexisNexis files, and perform data sanitization to remove any unwanted meta data. In particular, the parser extracts the title, body, source and publication date from the article, and disregards the remainder of the text. The text is then passed through an optional filter stage, which filters each text according to a set of user-specified keywords and a set of dates. In the Enron and pharmaceutical case studies I present in Chapter 4, I filter the corpus to ensure that each article contains one of a pre-specified set of firm names in the title, with at least three additional mentions in the body of the text. Both the set of firm names and minimum occurrence frequency are entirely customizable. The date filter allows the user to specify a set of filtering dates to exclude any articles outside of a particular time period. This allows the user to select a sub-sample of the corpus if he or she so desires.

²⁸ProQuest and LexisNexis are subscription-based content aggregation services, comprising news, magazines, journals, court documents, and many other types of documents.

²⁹Both services also allow the user to search for documents labelled with a specific company name, industry or ticker information but these are not consistent within a source, nor between different sources.

3.2.2 Pre-processing

Having chosen the textual data to use as input to the system, the next stage is the pre-processing pipeline. The pre-processing pipeline aims to transform the data into a format that can be used throughout the rest of the system. At the heart of this is the Token object, which stores both the string representation of a given word and a list of any General Inquirer tags that have assigned to it³⁰. The purpose of the pre-processing stage is therefore to convert each document into an array of Token objects, which are then used during the sentiment extraction phase.

Table 3.1: UML diagram for the Token class showing core data members.

Word
wordCount : int
senseID : int
fullWord : String
stemmedWord : String
tags : List(Integer)

The core structure of the Token object is shown in the UML class diagram in Figure 3.1. Each Token can be either a single word or a multi-word expression (e.g. named entities or idioms), and the wordCount field in the Token object keeps a count of the number of individual words in the entire string representation. The senseID field is used to store the corresponding sense number for ambiguous words in the GI dictionary, or is set to 0 in the case of unambiguous words³¹. The Token’s tags list is used to store any GI categories that are later assigned during the lookup/disambiguation phases. Each Token object contains two string representations: the original string representation (the variable ‘fullWord’) and, if applicable, a stemmed string representation (the variable ‘stemmedWord’). The system first performs a lookup using the original string

³⁰This includes any of the 183 semantic tags, as well as any marker tags (such as for punctuation tags or suffix tags) that are used for disambiguation.

³¹Words that have yet to be resolved (either directly via the lookup procedure, or else via the disambiguation routines in the case of ambiguous words) are assigned a senseID of -1, indicating that the word has yet to be processed.

representation and, upon encountering a negative result, performs a second lookup using the stemmed representation. The actual stemming process and lookup procedure will be described in greater detail in the sentiment extraction stage.

Thus each document is represented as a two-dimensional array of Tokens, where the first dimension corresponds to sentences within the document and the second dimension corresponds to individual Token objects. The purpose of the pre-processing stage is to convert each unstructured text into this array of Token objects. To this end, the pre-processing pipeline undergoes four main tasks:

- Duplication detection where duplicate documents are removed from the input corpus.
- Sentence splitting in which each document is broken down into individual sentences.
- Tokenization in which each sentence is broken down into individual Token objects.
- Named Entity recognition where any of the named entity Token objects are labelled as such to prevent further processing.

The duplicate detection routine is applied to the entire input corpus upon running the system. The sentence splitting, tokenization and named entity recognition tasks are applied to each individual document within the corpus. These latter tasks are highly dependent upon one another, and a single unified toolkit - the Stanford Core NLP toolkit (Manning et al. (2014)) - is used to perform each of these three tasks in parallel. In the next few sub-sections I describe these pre-processing components in more detail.

3.2.2.1 Duplicate Detection

The issue of duplicate documents arose during the case study stage, where I found that an article written about two companies of interest would be returned twice when searching for each of the companies' names. This was particularly common in the pharmaceutical case study, where articles were often written about multiple competitor companies within the same industry. In addition, it was also evident that some news providers would disseminate news across multiple channels. For instance, both the Wall Street Journal and Financial Times newspapers have a printed edition as well as an online edition, and occasionally they re-print the information across both editions

(sometimes on different days). Furthermore, some news vendors such as PR Newswire or Business Wire circulate texts across their local, national and international wire services, leading to multiple copies of the same document. One might argue that if two duplicates are published on the same day, then their sentiment counts would average out. However, the obvious corollary argument to this is that the effect of duplication could potentially obscure other events that appear on the same day. For example, if there was one negative article about a company j on day t , but three duplicate positive stories about a company i on the same day t , then the news about company i would obscure the news about company j as the frequency counts are computed at the aggregate level³². The presence of duplicate documents may therefore obscure the true effect of the sentiment.

To remove duplicate documents I created a duplicate document removal routine, which compares the byte content of each document in the input corpus³³, and deletes the exact duplicates from disk. In the event that the duplicates occurred on different days, I simply kept the earliest of the duplicate files and discarded the rest. The motivation behind this was that the older texts were likely to contain original information, and would therefore have more of an impact upon financial markets than newer stories. Although this method of duplicate detection was somewhat crude it worked remarkably well, removing almost 3,000 duplicate files from the pharmaceutical corpus. One possible area for future work would be to detect and remove near-duplicates, using a method such as shingling. Shingling involves partitioning each document into a set of ‘shingles’ (n-grams), and then comparing the number of shingles in common for a given document pair. Documents with a high overlap in shingle sets will be more similar than those with a lower overlap.

3.2.2.2 Sentence Splitting

The sentiment analysis system works at the sentence level, processing each document on a sentence-by-sentence basis. This is predominantly due to the way in which the GI disambiguation rules operate. The disambiguation rules operate by examining the

³²I concatenate all articles on a given day into one large article during the sentiment extraction stage.

³³A more computationally efficient solution would be to compare the hashcode of each file, however there is a chance that this could lead to collisions.

context in which a word appears: each rule checks for the presence (absence) of a specific set of words or tags, and will change the selected word sense according to what is found. For instance, one of the rules for the entry ‘fell’ specifies a search for the words ‘in love’ in the two words that follow. If ‘fell’ occurs as part of the phrase ‘fell in love’, then the entire sequence is treated as an idiom, and the categories for sense 8 of the word ‘fell’ in the GI dictionary are applied. However, if the word ‘fell’ is followed by the word ‘asleep’ (i.e. as part of the idiomatic phrase ‘fell asleep’), then the content categories for sense 9 are applied.

The rule context window is typically just one or two words to the left or right of a word, however sometimes longer-range dependencies can exist. Occasionally, the rules specify to search from the start of (or until the end of) a sentence, however the context window does not cross sentence-boundaries: the authors in Kelly & Stone (1975) believed that most contextual information required for disambiguation could be found intrasententially. As a result, each sentence is treated independently of every other sentence in the document. The purpose of the sentence splitting stage is therefore to split each textual document into an array of sentences such that each sentence can be processed independently from the last.

The problem of sentence splitting is itself a challenging task. Simply splitting a document based on end of sentence punctuation marks (e.g. ‘.’, ‘?’, ‘!’) is insufficient, as this ignores punctuation-containing words such as numbers, acronyms and abbreviations. Choosing words that start with capital letters as a sentence boundary is also insufficient, as proper nouns, acronyms and other initialisms can occur in the middle of a sentence. It is unclear how the authors of the original General Inquirer system performed this tokenization task and, unfortunately, its chief architect passed away several years ago. Rather than using or developing my own set of subjective heuristics, I decided to use a pre-existing API to perform the sentence splitting process. Specifically, I used the Stanford Core NLP toolkit (Manning et al. (2014)), which uses a deterministic tokenizer to mimic the Penn Treebank tokenization process. Intuitively, the Stanford Toolkit is also capable of performing word tokenization and named entity recognition (amongst other things) at the same time as splitting a text into sentences. Given its accuracy, speed and the ability to perform these three tokenization tasks in a single pass, it provides a logical choice for use in this system.

3.2.2.3 Named Entity Recognition

During the development of the system it transpired that named entities could be a potential source of ambiguity. For instance, while the word 'united' might be evaluative in certain contexts it would not necessarily be evaluative when used as a named entity, such as in the country names 'United Kingdom' and 'United States', the company name 'United Health' or the sports team 'Manchester United'. While the usages of evaluative terms within named entities could arguably be considered edge cases, distinguishing between named entity and non-named entity usages is even important for non-evaluative terms. For instance, consider the distinction between the word 'Apple' when used as a company versus its other common usage as a fruit. While neither of these senses are evaluative, it is important to distinguish between the two senses as they could potentially lead to the assignment of different semantic tags from the GI dictionary. These semantic tags are subsequently used in the GI's disambiguation routines, and so applying incorrect semantic tags could potentially lead to erroneous disambiguation of other words.

The problem is further compounded when using the stemming component of the system to reduce inflected words to their stemmed form. During the stemming process the system applies certain stemming tags to words, which are later used by the GI's disambiguation rules. For instance, words which are stripped of an '-er' inflectional suffix are assigned an ER marker tag, and words which are stripped of an '-s' inflectional suffix are assigned an S marker tag. These marker tags are subsequently used in the disambiguation routines for other words. Naturally, this can lead to undesired behaviour if applied to named entities. As a practical example from this thesis, consider the company 'Eli Lilly' which was one of the companies used as a case study Chapter 4. The stemming algorithm used in this system reduces 'Lilly' to the root form 'Lil' and applies the LY marker tag. If another word in the sentence is trying to distinguish between its noun and verb form by looking for the LY tag (which is typically used to mark adverbs), then the program might accidentally assign the wrong word sense.

To overcome the aforementioned problems, named entities are flagged by the system to prevent the accidental assignment of tags from the GI dictionary. The system checks for named entities during the tokenization phase, using the Stanford Core NLP toolkit (Manning et al. (2014)). Once an entity is recognized (which may span more than one word), the system marks it as 'resolved' in the same manner as idiomatic expressions,

preventing it from being further processed.

3.2.2.4 Tokenization

During the pre-processing stage the sentence is also tokenized into words using the Stanford Core NLP Toolkit. However, merely tokenizing words according to the Stanford tokenizer is inadequate for a number of reasons. The GI disambiguation rules rely upon certain words being tagged with marker tags that are used as part of the disambiguation procedures. For example, punctuation characters (e.g. ‘.’, ‘,’, ‘!’, ‘?’, ‘-’) should be tagged with the PUNC marker tag, which is used in a number of disambiguation rules. The first and last non-punctuation words in the sentence should also be tagged with the categories B and E, marking the beginning and end of the sentence respectively. During the tokenization phase contractions such as ‘I’ve’ and ‘we’ll’ are reduced to their root forms - in this case the pronouns ‘I’ and ‘we’. Any marker tags applicable to the removed clitics are subsequently applied to the root word to which the clitics were attached. For instance, Kelly & Stone (1975) state that root word to which the clitic ‘-ve’ is attached, should be tagged with the GI marker categories of HAV, VERB and SUPV. These clitic marker categories may subsequently be used in the disambiguation routines of other words.

The GI system also designates marker tags to tag numbers with their appropriate tags. In addition to the NUMB tag, there is also a CARD tag for cardinal numbers and an ORD tag for ordinal numbers. While all of the ordinals, and some of the cardinals appear as individual entries in the GI dictionary, the integer representations (e.g. 35, 1968) do not. These numbers are particularly common in financial news texts, which frequently discuss share prices and other monetary amounts. To overcome this issue, I use a regular expression to check if each token is a number. If it is, I apply the NUMB and CARD tags to every individual integer token.

The tokenization process then is naturally more complicated than simply following the Stanford Tokenizer’s verbatim tokenization procedures. The main point to note is that certain marker tags are inserted during tokenization that are used as part of the GI’s disambiguation procedures. In addition to special marker tags for numbers and punctuation marks, marker tags are also applied to the root word when removing enclitics. Named entities are also treated differently from non-named entities, to prevent the system from incorrectly applying sentiment tags to polysemous tokens.

3.2.3 Sentiment Extraction

The previous section described the pre-processing stage, which is applied to each of the documents in the input corpus. The outcome of this stage is a two dimensional array for each document, in which the first dimension denotes the sentences within the document, and the second denotes the Token objects. Note that at this point in time some of the Token objects have already had certain tags applied, such as punctuation markers, sentence boundaries, clitics, entities and others. Additionally, some of the Tokens have already been ‘resolved’ to the correct (unambiguous) sense, such that further processing of these Tokens is not required. The next stage is the sentiment extraction or ‘tagging’ phase, and this is where the dictionary lookup and disambiguation routines are applied, and tags are assigned to each of the individual Token objects. The output of this phase is a document x category frequency matrix denoting the counts of the number of ethical and evaluative terms for each of the input documents.

In addition to the aforementioned Token array, the sentiment extraction phase takes several lexical resources as input. The first of these is the GI dictionary, in which word senses are annotated with their applicable semantic tags. While in theory any other sentiment dictionary could be used instead, the work in this thesis solely uses the GI dictionary. Any other sentiment dictionaries would need to be monosemous (comprising of single senses) in order to be used ‘as is’ in the current system. The second major resource taken as input at this stage is the GI’s list of disambiguation rules, which define how to resolve ambiguous terms to their correct sense in the dictionary. These rules are naturally confined to the GI dictionary, and would not work with any other sentiment dictionaries. The third and final lexical resource used as input in the sentiment extraction phase is an optional domain ‘override’ dictionary. The override dictionary lists any domain-specific vocabulary which does not typically carry evaluative connotations in the target domain. As a result, any words in the override dictionary will be skipped by the system to ensure that they are not mistakenly identified as evaluative.

The sentiment extraction phase consists of four interlinked components, that are applied to each document in the corpus:

- Lookup procedure: which checks if a given term is in the GI (or optional override) dictionary.

- Stemming: which reduces inflected words to their base form, and applies any applicable stemming marker tags to be used for disambiguation.
- Disambiguation: which disambiguates ambiguous terms to their correct sense.
- Document aggregation: which groups documents on similar days and outputs a sentiment time series.

The first three of these processes are tightly interwoven, and are executed concurrently for each Token in the Token array. The last two processes are applied to the entire Token array once each Token has been resolved to its appropriate sense. The remainder of this section describes these four processes in greater detail.

3.2.3.1 Lookup & Stemming

The system's lookup procedure is the most fundamental part of the system's architecture. I have already discussed the Token object and associated UML class diagram during the pre-processing stage (see Figure 3.1) and I now refer back to that here. The lookup procedure encompasses both the lookup and stemming tasks that were mentioned in Figure 3.2.

The system processes each sentence on a token-by-token basis, performing the following tasks for each word. If the user specified to use a domain override dictionary then the system first performs an initial lookup in this dictionary. If the entry is found in the override dictionary, the system sets the token's 'resolved' flag to true, and proceeds to the next token in the sentence. Note that in some cases the override dictionary can contain multi-word expressions, and so in the event of a multi-word entry the system first checks whether the headword (the first word in the multi-word sequence) occurs in the override dictionary and, if so, whether any of the subsequent words in the phrase occur in the dictionary. If all words in the multi-word expression occur in order, then the system treats the entire sequence as an idiom and sets the resolved flag to true to prevent further processing. The program then skips to the next word in the sentence following the multi-word expression.

If the word is not in the override dictionary (or no override dictionary is specified) then the system performs a lookup in the GI dictionary. If the word is not in the GI dictionary, then the word is stemmed and a second lookup is attempted. If at this point the word is still not found, then the word's resolved flag is set to true, no tags

are added, and control passes to the next word in the sentence. However, if the word (or stemmed word) is found in the GI dictionary then at this stage the word can be either monosemous (i.e. unambiguous) or polysemous (i.e. ambiguous). If the word is monosemous, then the system sets the word's resolved flag to true, and applies any tag specific to this single sense. If the word is polysemous, then the system's disambiguation routine is activated.

At each lookup stage (in both the override dictionary and the original GI dictionary) the system checks whether both the full word or its stemmed equivalent are in the requisite dictionary. For instance, while the full word 'loved' is not in the GI dictionary, the canonical form 'love' is present. Pseudocode for the stemming algorithm is given in the appendix of this thesis.

The purpose of the override dictionary is designed to overcome shortcomings with adapting a general language dictionary to a particular domain. Specifically, certain lexical items are prone to take a different meaning in one domain compared with another. For instance, in the pharmaceutical case study to be presented in Chapter 4 certain negative and vice terms in the GI dictionary would not typically be considered negative in the pharmaceutical domain. The words 'cancer' and 'symptom' for example are tagged as negative in the GI dictionary, however these words appear relatively frequently in the pharmaceutical corpus in a non-evaluative sense. Consequently, the override dictionary allows one to ignore the GI dictionary if so required.

3.2.3.2 Disambiguation

The disambiguation procedures involve multiple iterations over each sentence in the document. For each word in a given sentence the system performs a lookup in the GI's disambiguation rule table. This is essentially a mapping structure that maps an ambiguous word onto a set of all disambiguation rules. The rules are executed sequentially in order as far as possible. Each rule requires examining the context around a word to check for the presence (absence) of other words or tags. If a rule is able to resolve a word to its correct sense then the senseID is set, the word's resolved flag is set to true, and any applicable tags are added to the word's tag list. Sometimes a word cannot be resolved during the first pass, as one of the disambiguation rules is checking for the presence of a particular GI tag later in the sentence which has yet to be applied. In this case resolution of the given word is deferred until subsequent iterations

over the target sentence, and control passes to the next word in the sentence.

If all words have been successfully resolved to their correct (most likely) senses, then the sentence is marked as complete and control is passed to the next sentence in the document. However, if there are ambiguous terms which have yet to be resolved, then subsequent passes are required. For instance, one of the disambiguation rules for the word ‘bank’ involves looking for the presence or absence of the GI’s ECON (economic) tag in the 5 words preceding or 5 words following it. During the first pass the ECON tag will not have been applied to any words following ‘bank’, as these words have not yet been processed by the system. During the second pass (or subsequent passes) the ECON tag will be applied to all words following ‘bank’, and so the system will be able to correctly determine whether the success or failure branch of this rule should be followed. However, even with a second pass the system might still be unable to completely resolve the words following ‘bank’, as these words might themselves be ambiguous. For instance the word ‘account’ (as in the phrase ‘bank account’) has 8 senses in the GI dictionary, only 2 of which are labelled with the ECON tag. In order to successfully resolve ‘bank’ to its correct sense, it may first be necessary to resolve the word ‘account’³⁴. In this case (and in the event that none of the other rules are able to correctly resolve the word), the processing of ‘bank’ will be deferred until a later pass, after ‘account’ (and any of the other words following ‘bank’) have been resolved.

However, even with subsequent passes it might be the case that the system is unable to completely resolve a sentence as two words might have disambiguation rules which are dependent upon the successful disambiguation of the other word. For instance, in the phrase ‘just like’, both ‘just’ and ‘like’ are ambiguous, and have disambiguation rules which depend on the following/preceding word respectively. In this case the GI’s ‘breaklock’ routine is activated (Kelly & Stone (1975)). Briefly, this breaklock routine involves considering whether the majority of ambiguous senses for a given word contain the tag that is being searched for, with an acceptance criteria of at least 80%. For instance, assuming that ‘bank’ and ‘account’ had dependencies upon one another, then the breaklock routine would involve checking whether at least 80% of the senses of ‘account’ contain the tag ECON: if so, the rule would fire, a match would be made, and the success branch would have been taken for the word ‘bank’. The system can

³⁴This may not be necessary if one of the other disambiguation rules for ‘bank’ has successfully resolved it to its most likely sense

then proceed to correctly apply the remaining rules to the word ‘account’³⁵.

One of the issues which this thesis does not directly address is that of contextual valence shifters - syntactic structures which can alter the intensity or direction of sentiment bearing phrases. Intensifiers such as ‘very’ or ‘exceptionally’, for instance, can potentially increase the potency of sentiment as in the phrase ‘very good’. Contrariwise, diminishers such as ‘slightly’ or ‘rather’ can reduce the potency of sentiment, as is the case in the phrase ‘slightly good’. A significant set of valence shifters is that of negators such as ‘not’, ‘nothing’ and ‘never’, which can reverse the entire polarity of a phrase. While there has been some effort to explore the roles of contextual valence shifters in studies on sentiment analysis (e.g. Pang et al. (2002), Polanyi & Zaenen (2006), Kennedy & Inkpen (2006)) many of these are confined to the context of classifying the orientation of reviews rather than financial sentiment analysis. Two exceptions which studied the role of negation in finance are Das & Chen (2007) and Jegadeesh & Wu (2013). Loughran & McDonald (2016) argue one potential reason for the lack of interest in negation in sentiment analysis studies in finance. In their paper they claim that while positive words are frequently affected by negation (e.g. ‘*not good*’), negative words tend not to be negated in the same fashion. Given, they claim, that most of the previous research has focused on negative rather than positive sentiment, the issue of negation has not been well explored.

There are several reasons for excluding the study of contextual valence shifters in this thesis. The first is that in this thesis I adopted a uniform weighting scheme whereby each term is equally important in the text³⁶. Allowing for intensifiers, diminishers and negators would have necessitated a deviation from this weighting schema. The assumption of a uniform weighting scheme was based on the confines of the GI dictionary. Although an earlier version of the GI dictionary had different categories for different strengths of positive and negative terms, these were subsequently combined into a single category in a later revision, as the authors believed it was too difficult for the annotators to agree upon the polarity strengths. A second reason for choosing not to allow for valence shifting structures was because there is very little agreement

³⁵Kelly & Stone (1975) discusses a few other subtleties to this breaklock routine. While these subtleties are applied to the system, they are not repeated here as the breaklock routine is rarely activated. The interested reader should consider Kelly & Stone (1975) for a list of the complete breaklock routine.

³⁶The legitimacy of this assumption is left as a potential area for future research.

upon how to accomplish this. The issue of negation scope, for instance, is a particular source of disagreement: Jegadeesh & Wu (2013) argue that only sentiment-bearing words within 3 words of a negator should be negated, whereas Das & Chen (2007) argue that all sentiment-bearing words up to the next punctuation marker should be negated. Attempting to incorporate this would mean explicitly defining the notion of negation scope, which is not a trivial task in itself. A further source of disagreement on valence shifters in the literature is how to weight the newly shifted valence. Hogenboom et al. (2011) suggests that reversal of sentiment polarity will not necessarily have the same strength as the alternative antonym. For instance, the authors refer to the word ‘bad’, and state that the negated counterpart ‘not bad’ is not as strong as the antonym ‘good’ and should therefore not be given the same sentiment strength. The third reason for not explicitly handling valence shifters in this thesis is that because it has been so rare in the literature on sentiment analysis in finance and accounting, it would not be easy to compare the results with existing work. Overall, excluding the handling of contextual valence shifters reduces the complexity and subjectivity of the system, however a thorough study of these syntactic markers and their impact in determining the sentiment in financial news is deferred as an idea for future work.

3.2.4 Document Aggregation

In the aggregation stage the system takes as input a list of all parsed documents and their associated dates. Each document is a two dimensional array of Token objects, where the first dimension represents sentences in the document and the second dimension represents the position of the Token in the sentence. Although this array is ordered, the word order here now becomes irrelevant and the job of the aggregation stage is to count the number of negative, vice and total words occurring in each document. The aggregation stage also groups documents with similar dates. For instance, if working with daily data the aggregation stage treats all documents occurring on the same day as an individual document. Thus for a given day ‘t’, the system provides raw frequency counts of $negative_t$, $vice_t$ and $total_t$ where the negative and vice variables represent counts of the negative and vice terms in the GI dictionary³⁷, and the total variable

³⁷Note that while some studies have found a relationship between positive terms and financial returns, most have found that the relationship is stronger for negative terms.

represent counts of all words in the document (note that multi-word expressions such as idioms and certain named entities are counted as the number of individual words in the phrase). The counts are subsequently normalized by dividing the raw frequencies (e.g. of negative and vice terms) by the total number of terms on a given day ‘t’.

The output of the aggregation stage is a frequency matrix, where rows correspond to article dates (aggregated by day, week, month etc.) and columns correspond to dictionary categories such as negative or vice. The cells in this matrix denote the relative frequency of a dictionary category at a given point in time. The frequency matrix is written to disk as a .csv file, which is subsequently used as input to the statistical modelling pipeline. At this stage the text analysis pipeline is complete, and the next stage of the system is to align the dates of the sentiment time series with the dates of the transformed market data.

3.3 Model Estimation and Evaluation

The statistical modelling component takes the sentiment time series output from the textual analysis component as input, together with relevant market data. The market data is then transformed to ensure that it meets the requisite criteria for performing the statistical tests, and then subsequently aligned with the sentiment data. The aligned time series is then input to a model estimation framework, in which summary statistics are calculated, and either a VAR or a rolling VAR model is computed to describe the interrelationships between the sentiment and market data. The output of this process is a set of model statistics that enable inferences to be made about the interrelations between sentiment and financial returns. An overview of the statistical modelling pipeline is given in Figure 3.3.

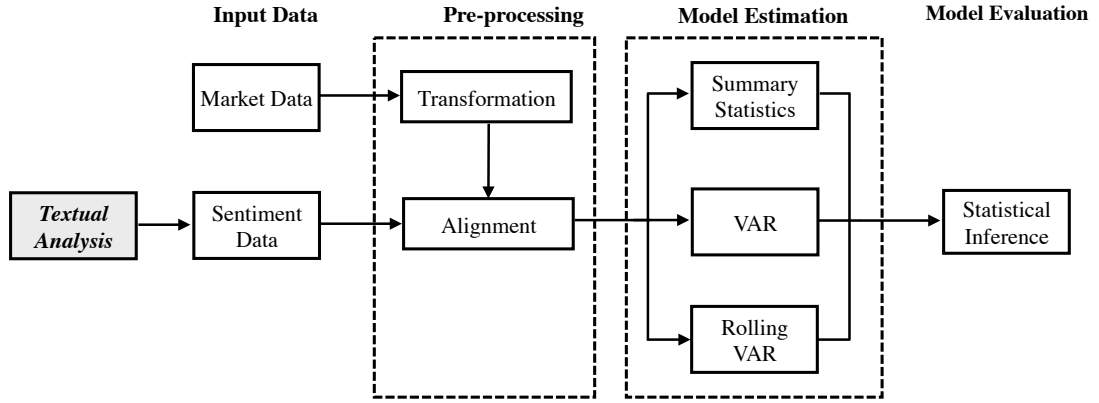
The statistical modelling pipeline essentially consists of four main components:

1. A transformation component where market data is transformed into an input that meets the model’s assumptions.
2. An alignment component which aligns the transformed market data and sentiment time series produced in the text analysis stage.
3. A model estimation component, which computes summary statistics, vector autoregression models, and rolling regression models if desired.

4. A statistical inference component, which outputs the results of the regression model describing the relationship between sentiment and returns.

These components are dependent upon one another, in that the sense that the output of each component is used as input to the next component. The rest of this chapter discusses each of these components in turn.

Figure 3.3: Key components of the statistical modelling pipeline designed in this thesis.



3.3.1 Market Data

In this thesis it is assumed that there is a linear relationship between share prices and sentiment, and that this relationship can be modelled using a vector autoregression (VAR) framework. As input, the statistical model requires market data in the form of share prices, together with any other market variables to be incorporated in the model. In this thesis I also include volume and volatility variables, which are known from the literature to have some impact on financial returns.

Historically, studies on sentiment analysis have tended to focus on prediction models that explain changes in asset prices at either the index-level (using a market index such as the Dow Jones Industrial Average (DJIA)), or at the individual stock level (using firm-level share prices)³⁸. In both cases the historical price data is represented as a time series - ‘a set of observations that a variable takes at different times, such as daily... weekly... monthly... quarterly... annually... quinquennially... decennially...’

³⁸Some studies have used other financial variables such as earnings information, however these are not discussed here as they are beyond the scope of this thesis.

(Damodar (1999)). A time series thus represents a variable that is repeatedly measured at equally spaced points in time. In this thesis, I focus on the use of *daily* asset prices, where each observation in the price time series corresponds to a given trading day. However, the system presented in this thesis is also capable of working with weekly or monthly data, and could easily be extended to other time periods if required. In the remainder of this section I discuss the main market data that is used as input to this sentiment system, together with any transformations of this data that may be required.

3.3.1.1 Returns

The main aim of this thesis is to determine the effect of textual sentiment on asset returns - the change in the price of an asset at a given point in time. In the context of investing this means the amount an asset increases or decreases in value relative to its original cost. Because returns are a proportional measure, they allow for easy comparison of financial instruments of different sizes, different currencies and across different markets, which would be difficult for one to accomplish using standard stock prices. One of the primary reasons for using returns instead of prices is that raw prices tend to exhibit heavy autocorrelation³⁹. This can cause problems for some statistical methods that require observations (or the error term) to be independent. Conversely, returns are generally not autocorrelated - a phenomenon that has been shown to be pervasive across various markets, securities, industries and countries (Taylor (2011)).

Financial returns also possess a number of other attractive statistical properties - known in the literature as *stylized facts* - that have been shown to be universal across different securities and markets (Taylor (2011) and Cont (2001)). For instance, returns are generally non-normally distributed, exhibiting more peakedness and fatter tails than would be found in a normal distribution. Despite their non-normal behaviour, the distribution of returns is generally symmetrical, and exhibits mean reversion behaviour whereby in the long-run returns tend to reverse to their mean value. Furthermore, returns sometimes exhibit volatility clustering: periods in which large returns (positive or negative) are followed by similarly large returns (positive or negative), and small returns are followed by small returns (Mandelbrot (1963)).

³⁹The term ‘autocorrelation’ refers to the correlation of observations from a time series with its past (historic) observations.

Although there are several different formalisms for calculating returns, the most common definition - and the one used in this thesis - is that of *continuously compounded returns*, which is defined in Equation 3.1.

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \quad (3.1)$$

Where:

r_t = the return at time t

p_t = the closing price at time t

$\ln$ = the natural logarithm

The choice of t could be any continuous period of time, such as daily, weekly or monthly. However, in this study, I work solely with daily data so treat r_t as the return of an asset on a given day t . Thus the return at time t represents the change in price between day t and day $t - 1$. The purpose of the logarithmic transformation is to ensure that the variance in the returns series remains constant, as a non-constant variance in the series can lead to spurious regression results. However, an attractive by-product of this transformation is the fact that returns then become time-additive: from daily returns it is possible to calculate weekly returns and so on. Another attractive property of this transformation is that when the difference between p_t and p_{t-1} is small, the logarithmic difference approximates a percentage change.

The system uses the *zoo* package in R to read historical share prices from a tab-delimited file. In this thesis all share price information is adjusted for stock splits and dividends by the data provider from which it is obtained⁴⁰. The system transforms the raw prices into returns using Equation 3.1, and stores the result in a time series object that is used throughout the rest of the system.

3.3.1.2 Measures of Market Liquidity and Risk

In this thesis asset returns are modelled as a function not only of historical returns and sentiment, but also of other market variables. These additional market variables have

⁴⁰This adjustment of share prices for stock splits and dividends is the default behaviour of most major financial data providers (e.g. Yahoo Finance, Google Finance, Quandl), and removes gaps or sudden spikes in the series caused by these corporate actions.

been shown in the literature (e.g. Tetlock (2007), García (2013) and Zhang et al. (2016)) to explain some of the variance in returns. Consequently, including them prevents the sentiment variable from accidentally capturing any variation attributable to other market variables.

The first of these additional explanatory market variables is volume, which acts as a proxy for market liquidity⁴¹. An asset's volume is essentially the number of shares traded over a given period (here, over a given day). A higher number of shares for a given stock suggests that the stock is more liquid, which causes investors to value it more highly (Aldridge (2013)). Additionally, more liquid stocks are said to incorporate information faster than illiquid stocks, which creates a potential for information arbitrage (ibid.). In Tetlock (2007), the author includes the volume of shares traded on the NYSE as an explanatory variable in his vector autoregression model. This ensures that the model does not unduly assign excessive weight to the sentiment variable, when in fact the variation in returns is actually caused by higher market liquidity. It is important to note that this volume variable is dynamic in that there are certain periods where volume is higher (lower) than other periods. This changing volume is modelled explicitly in the vector autoregression equations.

The second additional explanatory market variable included in the regressions is volatility: *'The extent to which a series is highly variable over time'* (Brooks (2014)). The volatility variable is essentially a measure of market risk, and measures the extent to which returns fluctuate around their average value. There are many different measures of volatility, including the standard deviation of returns, the log difference between the highest and lowest return at a given time t , or using squared daily returns (Brooks (2014)) (a variant of the latter is used in the market-level and firm-level case studies in Chapter 4). Most measures of volatility are derived from historical values of market prices, and can thus be calculated from the input share price information. However, the system is also capable of taking an external volatility proxy as input such as the Chicago Board Option Exchange's Volatility Index (used in the industry-level case study in Chapter 4). In general the choice of volatility proxy is usually unimportant, as the model results in Chapter 4 are robust to alternative volatility measures.

⁴¹The term liquidity here refers to the ability to buy or sell an asset without having to significantly alter its price. Therefore a more liquid asset is easier to sell than a less liquid asset.

3.3.2 Transformations

One of the assumptions of the vector autoregression model is that the time series used are all *stationary*, in that their means and variances remain constant over time. In time series data this is often not the case, as trends, business cycles and other market phenomenon can cause the mean or variance to fluctuate wildly. This is evident in Figure 3.4 which shows an example of a non-stationary series with both a non-stationary mean and a non-stationary variance.

One way of ensuring that the time series becomes stationary is to apply some form of statistical transformation. In this thesis I adopt a similar methodology to Campbell et al. (1993). I first take a logarithmic transformation, in order to stabilize the variance in the series, and then de-trend the transformed variable by subtracting a backward moving average for the past N observations. In line with Tetlock (2007), I set the value N to 60 observations, although the results in this thesis are robust to using other window sizes. Because the data used in this thesis comprises daily observations, this equates to subtracting a 60-day backwards moving average from the log-transformed variable. This method of transformation is applied to the volume and volatility variables discussed in the previous section. Note that the returns series has already been transformed by taking the log difference in prices.

In addition to the de-trended log transformation described above, I also standardize all continuous regression variables to z-scores. This includes both the market data (returns, volume and volatility time series), as well as the sentiment time series produced in the text analysis stage. The z-transformation takes the following form:

$$z = \frac{x - \bar{x}}{\sigma_x} \quad (3.2)$$

Where:

x = the variable to be transformed

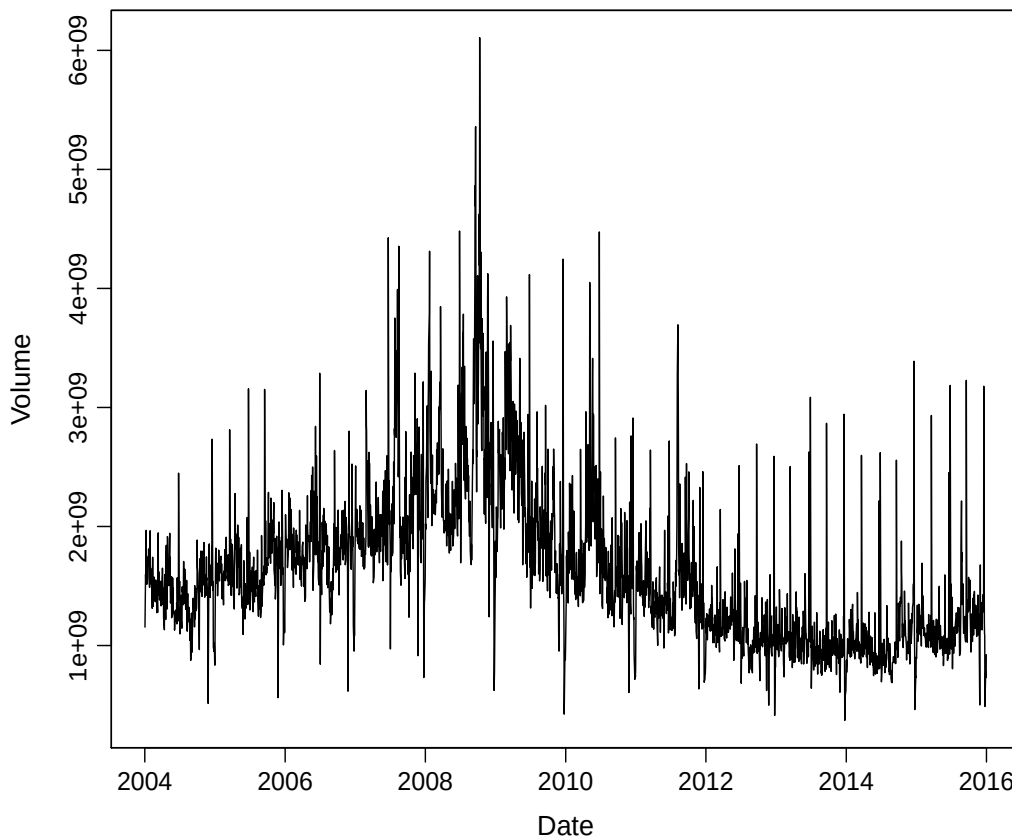
$\bar{x}$ = the mean of the variable x over the entire time series ($t_1 \dots t_N$)

σ_x = the standard deviation of the variable x over the entire time series ($t_1 \dots t_N$)

The purpose of this z-transformation is to ensure that all continuous variables are transformed to have the same units. The z-transformation results in a variable that is centred at zero and has a variance of one. As a result, transforming all continuous

variables to have the same units enables an easy comparison of the effect of one or more variables upon the other. Thus when comparing the bidirectional relationship between returns and sentiment (which have different original units), the strength of the relationship can easily be compared in terms of changes in standard deviation units.

Figure 3.4: Plot of NYSE Volume between 2004 and 2015. The data for this plot was downloaded from the NYSE Transactions, Statistics and Data Library (Available from <https://www.nyse.com/data/transactions-statistics-data-library>). Larger spikes correspond to larger variation, with particularly high variation during the recessionary period (2007-2009).



3.3.3 Data Alignment

Having transformed the market data into a form that meets the model's requirements, the next stage in the statistical modelling pipeline is to align the data from the financial time series with the dates of the sentiment time series produced from the textual analysis stage. While at first this seems like a straight-forward task, there are a number of practical concerns which need to be addressed.

Financial markets are open on every business day, and thus obtaining a continuous time series of market data is an uncomplicated task. Indeed a number of free websites

exist (e.g. Yahoo Finance, Google Finance, Quandl) that are devoted to this particular task. However, obtaining a continuous time series of news can be far more challenging. Even for the biggest companies it is often impossible to obtain news that occurs on every single business day, meaning that there could be gaps in the time series where there is no news. The question of how to deal with these gaps remains largely unresolved, as it is not clear what approach has been previously adopted in the literature. While this problem is not so severe in market-level and industry-level studies (which cover a wide range of companies and therefore have a much larger news coverage), it does present a complication in the study of individual firms, which may not have news written about them on every business day. One logical approach is to simply set days with no news to zero, reflecting the fact that no news equates to no sentiment on this particular day. This in fact is the approach adopted in the Enron case study, which I present in Chapter 4. Exploring the effect of alternative data alignment tasks remains an opportunity for future research.

Another problem with date alignment is that while financial markets are closed at weekends and on public holidays, news vendors often are not. Indeed, many newspapers (e.g. WSJ, Financial Times) even have special editions that are published during the weekend. This begs the question of what to do about this non-business day news. One approach would be to move weekend (or public holiday) news forward to the next available trading day. This would reflect the fact, for instance, that weekend information will be impounded into Monday's returns (Monday's returns are calculated as the log difference between Friday's and Monday's closing prices). However, on the other hand it is difficult to know what useful new information would be included in weekend news, as many major businesses would be closed during this period. In addition, many institutional investors tend to 'tune out' their investments over the weekend (DellaVigna & Pollet (2009)). Consequently, an alternative approach is to simply ignore weekend news (by treating the equivalent returns dates as NAs). This in fact is the approach adopted in this thesis, and indeed much of the existing literature follows a similar approach⁴². A third approach would be to perform some sort of interpolation of returns over the weekend, however this then introduces researcher subjectivity, as it begs the question

⁴²Note that in the regression models (to be discussed) I also include a dummy day-of-the-week variable for Monday, which is set to 1 on Mondays and 0 elsewhere. Any effect captured due to the weekend news will be incorporated into this variable.

of what interpolation method should be used. Finally, a fourth approach would be to use an alternative method of returns, such as the log difference between open and close prices during a given time period t , rather than the log difference between close prices during times t and $t-1$. This approach removes subjectivity from the equation, but then means that results will not necessarily be comparable with the existing literature which tends to use close-to-close prices to calculate returns. Deciding what to do with weekend data thus remains a potential area for future research.

3.3.4 Model Estimation

Having computed a sentiment time series using the system described in this chapter, and transformed a set of market variables using the statistical pipeline, the next stage in the system is to explore the effect of this sentiment proxy on a market phenomenon of interest. While in theory any market variable could be modelled in this thesis I am concerned primarily with the effect of sentiment on share returns. To this end I employ a vector autoregression methodology, which posits a linear relationship between sentiment and returns. Consequently, in this chapter I expand upon this discussion of vector autoregression models, paying particular attention to the model's assumptions and diagnostic tests.

3.3.4.1 VAR Model

The vector autoregression (VAR) model posits that a variable can be dependent both on its own past history, and on the history of other exogenous variables. Specifically, the model suggests that the relationship between the dependent and independent variables is linear, and can be represented in an additive fashion. Notationally, this can be represented as:

$$y_t = \beta_0 + \beta_1 L_n y_t + \beta_2 L_n x_t + \beta_3 L_n z_t + \varepsilon_t \quad (3.3)$$

Where:

y_t = the dependent variable at time t

x, y, z = the independent variables

β_0 = a constant term

$\beta_1 \dots \beta_3$ = the 3 model coefficients (weights of the independent variables)

ε = a random error term (assumed to have a mean of 0)

L_n = a lag operator where $L_n y_t = \sum_{i=1}^n y_{t-i}$

The lag operator L_n here is a notational convenience, defining a function which transforms a given variable into a linear combination of its historic n values. While in theory each independent variable could have its own unique lag structure, it is common in VAR models to assume the same number of lagged terms for each independent variable. In this study I set the lag structure to 5, such that the dependent variable will be dependent on its previous 5 values, and upon the previous 5 values of any other independent variables. Because the data used in this thesis comprises daily data, this lag structure of 5 equates to one prior trading week worth of values (it is also the value used in much of the previous literature including Tetlock (2007), García (2013) and Ahmad et al. (2016)). Using the previous 5 lagged values allows the researcher to investigate whether there is a delayed effect of a given variable. For example, it may be the case that there is a delayed impact between sentiment and returns, whereby the effect of sentiment is not immediately incorporated into the change in share price, but is incorporated over a number of days. It also allows the research to observe whether the impact of sentiment is transient or permanent.

In this thesis I am concerned with the relationship between negative sentiment, ethical terms and asset returns. In each of the case studies (presented in Chapter 4) I fit two regression models: one in which returns are regressed on negative terms, and one in which they are regressed on vice terms. The frequencies of negative and vice terms are determined by the text processing stage described earlier in this chapter. Therefore the essential components of interest in this thesis can be more aptly be represented in the following equation:

$$r_t = \beta_0 + \beta_1 L_5 r_t + \beta_2 L_5 \text{Sentiment}_t + \varepsilon_t \quad (3.4)$$

Where:

r_t = the returns at time t

Sentiment_t = the frequency of negative terms or vice terms

However, as previously mentioned there are a number of other market phenomenon that are known to have an impact on returns. Specifically, the variables of volume (which proxies for market liquidity) and volatility (which proxies for market risk) should also be incorporated into the model to prevent the sentiment variable from inadvertently capturing their effect.

It has also been observed in the literature (e.g. Taylor (2011) and Makridakis et al. (1998)) that a number of calendar anomalies exist, and that average equity returns vary according to these anomalies. For instance, one such calendar anomaly is the January effect, which stipulates that average returns in January are lower than in other months. The reason for this, it is claimed, is that investors sell off their losing securities towards the end of December in order to realize the tax losses. A similar phenomenon has also been observed on various days of the week (particularly those close to the weekend). To take account of these market anomalies, I include January and day-of-the-week dummy variables in all regression equations. The January dummy variable $January_t$ takes the value of 1 if the month t is in January, and the value of 0 for every other month. Similar dummy variables are introduced for the days of Monday, Tuesday, Wednesday and Thursday (although Friday is excluded to avoid issues of multicollinearity). Collectively, I define the notation $Calendar_t$ which denotes the set of dummy variables $January_t$, $Monday_t$, $Tuesday_t$, $Wednesday_t$ and $Thursday_t$.

Thus the overall VAR model fitted in this thesis, including both the volume and volatility variables, together with any calendar dummy variables, is of the form:

$$r_t = \beta_0 + \beta_1 L_5 r_t + \beta_2 L_5 Volume_t + \beta_3 L_5 Volatility_t + \beta_4 Calendar_t + \beta_5 L_5 Sentiment_t + \varepsilon_t \quad (3.5)$$

Where:

r_t = the returns at time t

β_0 = a constant denoting the mean of r_t when all other variables are zero

$Volume_t$ = the volume proxy for liquidity (de-trended and log transformed)

$Volatility_t$ = the volatility proxy for market risk (de-trended and log transformed)

$Sentiment_t$ = the frequency of negative or vice terms

$Calendar_t$ = the set of calendar dummy variables [$January_t$, $Monday_t$, $Tuesday_t$, $Wednesday_t$, $Thursday_t$]

3.3.4.2 Model Assumptions

Having defined the full VAR model that is used in this thesis (see Equation 3.5), I now turn to a discussion of the model assumptions. Although regression models have a large number of assumptions, there are two in particular that prove problematic for time series analysis:

- Stationarity: the mean and variance of the underlying series should be constant over time
- Uncorrelated residuals: the values of the residuals (error term) should not be correlated through time

Both assumptions should be met in order that the regression estimations are unbiased. Formally, statistics to test for these assumptions are applied automatically by the system, and are produced as output during the model evaluation stage.

3.3.4.2.1 Stationarity

Regression analysis using time series data assumes that the data is *stationary* (Damodar (1999)), in that both the mean and variance of the underlying data are approximately constant over time. Many of the statistical tests conducted using the estimated regression coefficients are directly tied to this assumption. For instance, both the t-tests and F-tests traditionally used in the evaluation of regression models assume that the data is homoscedastic. Violations of this assumption can lead to biased coefficient estimation, and ultimately to erroneous statistical inferences. In reality time series are often non-stationary as trends, business cycles, seasons and other market phenomenon can cause the mean or variance to change over time.

Often, it is possible to transform the original series, such that the time series becomes stationary. One such method is to first-difference the data, by subtracting the change in a series from one period to the next. Thus if x denotes a time series of N observations, for each x_t one would subtract the previous observation at x_{t-1} . This

was the approach adopted in the returns equation (Equation 3.1) which corresponds to the (log) first-difference of the price time series.

Another method for converting a non-stationary series into a stationary one is to de-trend the series using some form of smoothing technique. A common method used in this thesis and elsewhere in the literature (e.g. Campbell et al. (1993), Tetlock (2007)) is to subtract a backward moving average from the series. For instance, given a time series observation x_t , one could subtract the average of values from $\sum_{t-1}^{t-M}$ where M denotes the number of historical observations to be included in the average. A larger number of observations leads to a smoother series, however runs the risk of covering up genuinely useful patterns. In this thesis, following Tetlock's (2007) lead, I set the window size M to 60 days, however the observed results remain robust to using different window sizes.

Although there are various methods for determining stationarity, one of the most widely-used methods is the augmented Dickey-Fuller (ADF) test (Dickey & Fuller (1979)). The ADF test tests the null hypothesis of a unit root⁴³ by fitting the following autoregressive model:

$$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \sum_{i=1}^p \delta_i \Delta y_{t-i} + \varepsilon_t \quad (3.6)$$

Where:

β_t = denotes a deterministic time trend

α = the intercept term denoting a drift component

p = the number of lagged first-difference terms to be included in the autoregression

Δy_t = the differenced time series such that $\Delta y_t = y_t - y_{t-1}$

ε_t = a random error term (assumed to have a mean of 0 and a standard deviation of 1)

The model's coefficients are estimated from Equation 3.6 using the OLS procedure. The main focus of the ADF test is whether the coefficient γ equals 0. If the variable y_t does need differencing, then the estimated value of γ will be approximately 0. Oth-

⁴³The term 'unit root' refers to the square root of the coefficient γ in Equation 3.6. If the value of γ is equal to 1, then its square root will also be equal to 1, and the variable y_{t-1} will have useful explanatory power in determining Δy_t , implying y is non-stationary.

erwise, if the series is already stationary, the value of γ will be negative. Thus the ADF uses a t-test to test the null hypothesis that $\gamma = 0$, and that therefore the lagged value of the dependent variable (i.e. y_{t-1}) offers no useful information that enhances prediction of Δy_t . In other words, the null hypothesis implies that the series is non-stationary, and a rejection of the null hypothesis implies that the series is stationary. In calculating the ADF test statistic I use the `adf.test()` function in R's *tseries* package.

3.3.4.2.2 Uncorrelated Residuals

Regression analysis using time series data also assumes that the model's residual terms are not *autocorrelated*. The term autocorrelation refers to a time ordered variable that exhibits serial correlation between consecutive observations. In the present case the regression model assumes that there is no serial dependence between consecutive observations in the residual term ε_t . The implication of autocorrelated residuals is that statistical inferences made with the model's estimated coefficients may be biased. As a result, any statistical tests made using the regression coefficients (e.g. t-tests, F-tests, etc.) will be invalid, leading to questionable inferences.

Informally one could check for autocorrelation in the residuals using a graphical means, by plotting the residual values ε_t against their lagged residuals ε_{t-1} to see if any obvious pattern is evident. However, graphical patterns are often difficult to distinguish, especially with a large number of observations. Consequently, econometricians tend to turn to formal statistical tests in order to check for the presence of autocorrelation in the residuals. One common test, and the one used in this thesis, is the Durbin-Watson test for autocorrelation. The Durbin-Watson statistic tests the null hypothesis that the fitted model's residuals are not autocorrelated. The statistic is calculated using the following equation:

$$DW = \frac{\sum_{t=2}^T (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^T \varepsilon_t^2} \quad (3.7)$$

Where:

ε_t = estimated residual at time t

T = number of observations in the model

In Equation 3.7 the numerator calculates the sum of the squared differences between the residuals at time t and their lagged values at $t - 1$. The denominator denotes the residual sum of squares. In this thesis the Durbin-Watson test is administered using the `durbin.watson()` function from the *car* ('companion to regression') library in R. Results of the Durbin-Watson calculations for the case studies in this thesis are given in the appendix.

3.3.5 Model Evaluation

Having calculated the coefficients of the VAR model (usually using ordinary least squares or some other parameter estimation method) the next stage is to interpret these coefficients in terms of their statistical significance. In this thesis I make use of three primary diagnostic tools, that will be used extensively throughout Chapter 4:

- Individual coefficient tests: to test the impact of a single variable upon the dependent variable.
- Joint significance tests: to test the joint effect of two or more variable upon the dependent variable.
- Mean Square Error (MSE): to compare the overall accuracy of the model (or between models).

Any inferences made from the individual and joint significance tests assume that the estimated standard error term is non-biased. This means that the series should exhibit no heteroscedasticity (i.e. be stationary) and contain no autocorrelated residuals, as per the model assumptions previously described. However, heteroscedasticity-consistent (or 'robust') standard errors can be computed to account for heteroscedasticity and autocorrelation in the model. In this thesis, following Tetlock's (2007) lead, I use Newey-West standard errors to account for heteroscedasticity and autocorrelation up to five lags. Thus all statistics computed in this thesis, report Newey-West adjusted standard errors.

3.3.5.1 Individual Variable Tests

In testing whether an individual variable has a significant impact upon the dependent variable, this thesis makes extensive use of the t-test. The t-test compares whether an

estimated regression coefficient is significantly different from 0, positing the following null and alternative hypotheses:

$$\begin{aligned} H_0 : \beta &= 0 & (\text{null hypothesis}) \\ H_1 : \beta &\neq 0 & (\text{alternative hypothesis}) \end{aligned}$$

The test statistic for this hypothesis test is given by:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Where:

$\hat{\beta}$ = the estimated regression coefficient

$SE(\hat{\beta})$ = the standard error of the estimated coefficient

t = the test statistic which follows a t-distribution under H_0 with $n-1$ degrees of freedom

In this thesis, I perform t-tests on each lag of the negative and vice variables in equation 3.5, to investigate whether any of the historical sentiment proxies have an impact on financial returns. Any coefficient with a statistically significant t-value (assumed here to have an associated p -value of < 0.05) is considered to have a significant (non-zero) impact upon the dependent variable.

3.3.5.2 Joint Significance Tests

In addition to testing the significance of an individual regression coefficient, it is sometimes interesting to test whether a *group* of variables have a jointly significant impact upon the dependent variable. For instance in Equation 3.5, financial returns are posited as a linear combination of the previous 5 day's lags of a given sentiment proxy. While each of these lags could be tested individually (and independently of every other lag), it is also possible to test the joint impact of all five lags at once. Formally, this is posited as the following hypothesis test:

$$\begin{aligned} H_0 : & (\text{null hypothesis}) \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \dots \text{ and } \beta_n = 0 \\ H_1 : & (\text{alternative hypothesis}) \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \dots \text{ or } \beta_n \neq 0 \end{aligned}$$

In the null hypothesis case, all of the hypothesised variables are jointly equal to zero. In the alternative hypothesis case at least one of the coefficients is significantly different from zero. The joint hypothesis test may therefore be used to test whether all lags of the negative or ethical variables are simultaneously equal to zero, and therefore to examine their collective impact upon the dependent variable.

One method of testing for joint significance is to compare the results of two models - a restricted model in which the jointly hypothesised variables are omitted, and an unrestricted model in which the variables are included. This measure, known as the F-test, compares whether including the hypothesised variables explains any extra variability in the dependent variable that is not explained in the restricted model. In order to do so, the F-test compares the sum of squared residuals (SSR) in the restricted model to the sum of squared residuals in the unrestricted model:

$$F = \frac{(SSR_r - SSR_{ur})/m}{SSR_{ur}/(n-k-1)}$$

Where:

SSR = denotes sum of squared residuals

r = restricted model

ur = unrestricted model

m = number of variables in ur - number of variables in r

n = total number of observations

k = number of parameters estimated in the unrestricted model

Under the null hypothesis, the calculated F statistic is assumed to follow an F distribution with $(m, n-k-1)$ degrees of freedom. A large F value relative to the critical value (here $p < 0.05$) will result in rejection of the null hypothesis. The conclusion of this is that the hypothesised variables jointly impact upon the dependent variable.

Another method for testing the joint significance level is to perform a Wald test. As in the F-test, the Wald test tests the null hypothesis that a group of hypothesised regression coefficients are all simultaneously equal to zero. Under the null hypothesis the Wald test follows a chi-square distribution with degrees of freedom equal to the number of hypothesised variables. Although the Wald test and F-test are asymptotically

equivalent, the mathematics behind the test are quite involved. A detailed discussion of the mathematical proofs are available in Greene (2000).

In this thesis I use the Wald test to ensure comparability of my results with Tetlock (2007). Specifically, I test whether any of the 5 lags of negative or ethical variables have any jointly significant impact on asset returns. To do so I use the `linearHypothesis()` function, available in the *car* (companion to regression) package in R. However, the F-statistic above can be computed from the Wald statistic by multiplying the F-value by the number of hypothesised variables to be tested. In Chapter 4 I therefore use the Wald test to compute the joint hypothesis of the negative and vice variables. The Wald statistic will be referred to by its chi-square notation (χ^2).

3.3.5.3 Mean Square Error

The system presented in this chapter has various processing parameters, which all result in different model estimates. In order to compare the quality of different models, or select the most appropriate model from a set of potential candidates, some form of accuracy measure is required. One of the most commonly used statistical methods to measure the accuracy of a model is the mean square error (MSE). The MSE measures the difference between the values estimated by the fitted model, and the values that are actually observed by the model. More formally:

$$MSE = \frac{1}{T} \sum_{t=1}^T \varepsilon_t^2 \quad (3.8)$$

Where:

T = the number of observations in the model.

ε = the model's residual at time t .

Therefore the MSE represents the average squared error between the fitted and observed values. This is in fact an intuitive approach, given that the OLS algorithm works by minimizing the squared error term.

3.4 Summary

In this chapter I described the development of a system for constructing ethical and evaluative time series from unstructured texts. These time series can be subsequently used in econometric models to explore their relationship with financial returns. As input the system takes both textual and market data that are annotated with temporal information. This chapter described the procurement and sanitization of this data, together with required transformations. Ultimately the system performs the tasks of perusal, aggregation and inference, that are typically conducted by investors when reading the news.

At its core the system’s architecture consists of two main components: a textual analysis component and a statistical modelling component. The textual analysis component involves pre-processing and parsing the data and produces ethical and evaluative time series as output. The statistical modelling component takes the time series output from the textual analysis stage and explores its relationship with financial returns.

In the text analysis pipeline the most fundamental component is the extraction of sentiment from the pre-processed texts. For each word in a given input text a lookup procedure is performed. This involves searching for the word in the General Inquirer dictionary, and possibly stemming the terms to their canonical form. If a word is found in the GI dictionary then any of the word’s associated tags are applied. If not, then the word is stemmed and a second lookup is performed. However, it may be the case that the word is ambiguous in that it has multiple senses in the dictionary. In this case, the GI’s disambiguation procedures are activated. These disambiguation rules involve scanning the immediate context of the word and looking for the presence or absence of certain words or tags. Ultimately the word is resolved to a given sense, and tags associated with this sense are applied. Another feature of the sentiment extraction phase is the use of a domain-override dictionary. This dictionary comprises a list of domain-specific lexical items which are not typically evaluative in the target domain. If a word is found in this override dictionary then the system will ignore it and not apply any of the GI’s tags.

After pre-processing each of the input texts an aggregation operation is performed. This essentially involves aggregating documents of a similar date, and summing the counts of the negative and vice variables on similar dates. In addition the counts are

also normalized during this stage, by dividing the raw frequencies by the total number of terms on that date. The resulting ethical and evaluative time series are then passed to the statistical modelling stage, which explores the impact of the time series variables on returns. Typically this involves computing summary statistics, along with a VAR or rolling VAR model. To this end a number of modelling assumptions must also be met, and evaluation of the computed models must be performed.

The main premise of this thesis is to explore the interrelationships between ethical (evaluative) terms and returns. In the next chapter I explore this relationship in three main studies, using the system implementation described here. Each case study spans a different period of time with different market and textual data used. This allows for analysis of the system under different market conditions, as well as with various different input types. For each case study a series of VAR and rolling models are computed to describe the relationship between negative and vice terms and returns.

Chapter 4

Results

4.1 Introduction

In the preceding chapters the role of investor sentiment was discussed in the context of determining asset prices. It was contended that investors can be susceptible to emotional and psychological factors that can cause them to behave in ‘irrational’ ways at certain times. This ‘irrational’ behaviour can lead them to overvalue or undervalue a particular asset, pricing it above or below its ‘true’ (fundamental) value.

By analysing the texts typically read by investors, it is believed that a glimpse into investors’ thought processes can be gained. However, in an age where ‘Content is King’ the deluge of information can often be overwhelming. Perhaps the solution to this information overload is in the forms of intelligent systems, capable of rapidly digesting and acting upon the information contained in text. The previous chapter described the development of such a system, and in this chapter the system is evaluated using three case studies at the market-level, firm-level and industry-level. These case studies were chosen specifically to demonstrate the variety of textual and market data which the system is capable of dealing with.

Most of the existing work in sentiment analysis has focussed on the analysis of negative and/or positive language. It is believed that this evaluative language functions as a proxy for investor sentiment, with the presence (absence) of negative terms acting as a proxy for *bearish* (*bullish*) behaviour. In this chapter, I explore the notion of whether evaluative language is rooted in ethical perceptions. Specifically, I test whether a subset of the negative terms - the ones that carry unethical connotations - can be used to derive this investor sentiment proxy. The questions which I aim to answer are

as follows:

- Do vice terms impact market prices in the same manner as negative terms? Can the vice terms be substituted for negative terms, without any major loss of information?
- Is the impact of negative/vice terms confounded by domain-specific vocabulary?
- Is the impact of negative/vice terms stronger during certain periods?

To answer these questions three major case studies are presented, using financial data at the firm-level, sector-level and market-level, and textual data from a variety of sources. Each case study spans a different period of time, in order to explore the robustness of the results under different market conditions. For each study a number of VAR models are computed to show the impact of negative and vice terms on financial returns. Formal statistical tests are then estimated to test the impact of these sentiment proxies. A series of rolling regression models are subsequently conducted in order to evaluate whether the impact of negative or vice terms is confined only to particular time periods. The remainder of this section describes the structure of the rest of this chapter, together with an overview of each of the three case studies that will be presented.

The first case study seeks to validate the system developed in this thesis (see Chapter 3), by replicating and extending the results of a major previous study (namely that of Tetlock (2007)). In Tetlock (2007) the author explored the effect of negative sentiment at the market-level, using the Dow Jones Industrial Average (DJIA) index as a proxy for market value. A daily Wall Street Journal opinion column was used as the source of investor sentiment. By replicating this study and reproducing its core results I validate both the system's construction (in terms of extracting and constructing a negative sentiment proxy) and the statistical methods used. After replicating the major results of this study, the investigation is then extended to the use of ethical vice terms, using the same time frame, textual data and market data.

The second case study examines the impact of sentiment and ethical terms during times of financial misdemeanours. Specifically, the study involves an investigation into the Enron scandal, which resulted in bankruptcy of the company and led to an investigation by the US Security and Exchange Commission, ultimately culminating in the indictment of a number of former key executives. A corpus of firm-level news articles

were constructed from a variety of different sources, which include both editorial and opinion column news as well as more general reportage. The effect of sentiment and ethical terms is examined at the individual firm-level, using historical returns as a proxy for the value of the firm.

The third and final case study explores the impact of sentiment and ethical proxies in the pharmaceutical industry. The study derives sentiment and ethical proxies from a collection of news texts written about 25 US pharmaceutical companies. The impact of these proxies is then examined against returns of the New York Stock Exchange’s (NYSE’s) ARCA Pharmaceutical index. During this study the issue of domain-specific vocabulary is also addressed and its implication for constructing a sentiment proxy using a general sentiment dictionary.

4.2 Case Study 1: Sentiment at the Market-Level

4.2.1 Introduction

In this section the system outlined in Chapter 3 is evaluated by reproducing some of the key findings of Tetlock (2007). Tetlock was one of the earliest to explore the impact of sentiment in financial markets and showed that negative financial news written in a daily Wall Street Journal opinion column had a significant impact on next-day returns of the Dow Jones Industrial Average (DJIA) index. To compute his negative sentiment proxy, the author used the General Inquirer (GI) system, comprising both the aforementioned dictionary and its associated disambiguation rules. The system in this thesis also uses the GI dictionary and disambiguation rules, and so the produced results should be broadly similar to those observed by Tetlock. As a result, not only does this section provide independent confirmation of Tetlock’s results, but it also validates the construction of the system. In addition, this section also aims to validate the contribution of the GI’s disambiguation rules, by comparing results both with and without the disambiguation rules to investigate the potential effect of ambiguity on the sentiment proxy.

Following replication of Tetlock’s key results, the investigation is then extended to the GI’s ethical vice terms, which comprise a subset of the original negative terms. During these extended investigations, the same data, time period and method was used in order to allow for an accurate comparison between the sentiment and ethical proxies.

A series of rolling regressions is also performed in order to investigate the time-varying impact of both the sentiment and ethical proxies. These rolling regressions allow for the fact that the market may be efficient (inefficient) during certain time periods, and thus measure the suitability of the model parameters over time.

4.2.2 Data

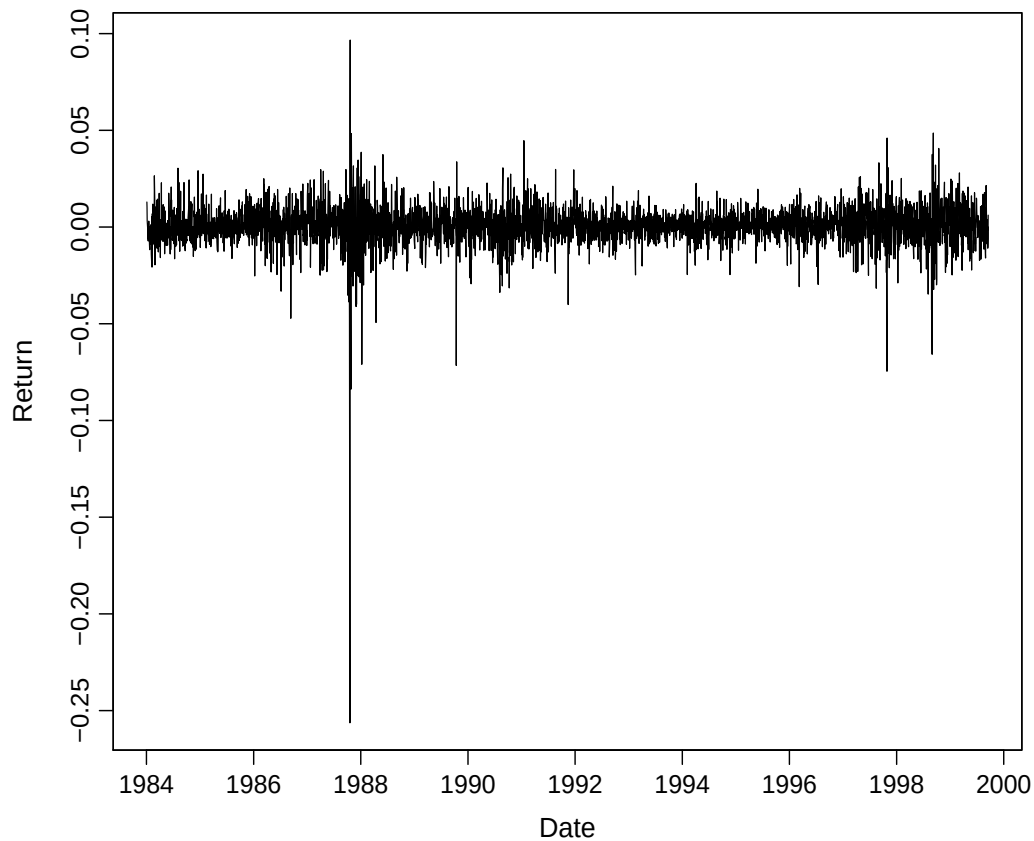
In Tetlock’s study, the Wall Street Journal’s ‘Abreast of the Market’ (AofM) column was used to explore the impact of news on next-day returns of the Dow Jones Industrial Average (DJIA) index. The AofM column was a daily opinion column representing a ‘post-mortem’ of the US stock market on the previous day. Typically, the column expressed how the markets changed on the previous day, outlining possible reasons for the changes and discussing the biggest movers of that day. While the column was primarily concerned with historical events, sometimes the column would also provide speculative statements about what the market might do next (Tetlock (2007)). This distinction is irrelevant for the purpose of this study, and historical and prospective information are treated equally.

Historical copies of the AofM articles were downloaded from the ProQuest news database, using the search criteria ‘COL(Abreast of the Market)’ and restricting the time period to 01/01/1984 - 17/09/1999 - the same time period used in Tetlock’s study. This resulted in a total of 3,963 daily texts. Each article was subsequently stripped of any unwanted meta-data using the system’s built-in ProQuest parser. There may, however, be some slight discrepancies between the meta-data removed in this thesis and the meta-data removed in Tetlock (2007). Because the column was published on business days, aligning the dates of the textual and market data was relatively straight forward.

As in Tetlock’s study, the Dow Jones Industrial Average (DJIA) index was used as a measure of market value. Historical DJIA prices were obtained from the Dow Jones Averages website (<http://www.djaverages.com>, Accessed: 16 July 2016), and subsequently transformed into continuously compounded returns in the manner previously described (see Section 3.3.1.1). In addition to lagged DJIA returns, the model also includes volume and volatility variables to capture other known sources of variation in returns. The volume variable comprised historic volume data of the New York Stock Exchange (NYSE), which was downloaded from the NYSE’s ‘Transactions,

Statistics and Data Library' (<https://www.nyse.com/data/transactions-statistics-data-library>, Accessed: 26 July 2016). The volatility variable was calculated directly from the returns variable, and will be discussed further in the next section.

Figure 4.1: DJIA returns over the period 01/01/1984 through 17/09/1999. The series shows that the average returns are centred around zero. The large dip of 19/10/1987 indicates the Black Monday Crash, where returns of the DJIA index fell by 22.61% on a single day, before subsequently reversing the following day.



4.2.3 Methods

The system processes the corpus on a document-by-document basis, splitting each document into sentences and each sentence into words, before performing a lookup of each word. For each document, the system keeps a record of the total number of negative and vice terms in each document, along with the total number of words in each document. The system then outputs a frequency matrix, where rows comprise individual dates and columns comprise total words, number of negative words and number of vice words. The matrix is subsequently exported to a .csv file, which is then imported into

the R statistical package for performing econometric analysis. Specifically, an R script written by the author automatically reads the frequency matrix into memory, normalizes the frequency counts, and fits the following regression equation as stipulated in Tetlock (2007):

$$DJIA_t = \beta_0 + \beta_1 L_5 DJIA_t + \beta_2 L_5 Volume_t + \beta_3 L_5 Volatility_t + \beta_4 Calendar_t + \beta_5 Crash_t + \beta_6 L_5 Negative_t + \varepsilon_t \quad (4.1)$$

Where:

$DJIA_t$ = the continuously compounded return of the DJIA index at time t

$Volume_t$ = volume of the NYSE to proxy for market volatility

$Volatility_t$ = volatility of DJIA returns to proxy for market risk

$Calendar_t$ = dummy calendar variables for month of January and days-of-the-week effects

$Negative_t$ = the standardized fraction of negative terms at time t

$Crash_t$ = a dummy variable denoting the Black Monday stock market crash of 19/10/1987

β_0 = a constant (denoting the mean of the returns when the other variables are set to 0)

ε_t = an error term assumed to have a mean of zero.

In Equation 4.1 L_5 defines the lag operator which turns a variable x_t into the row vector $[x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}]$, denoting its historic (lagged) values. The ‘Volume’ variable is log transformed to stabilize the variance, and subsequently detrended by subtracting a backward moving average for the previous 60 days. This transformation and detrending procedure is the same one used by Tetlock (2007), enabling comparison with his results and ensuring the model’s variables meet the assumptions outlined in the previous chapter. The ‘Volatility’ is constructed in the same manner described in Tetlock (2007): first the DJIA variable is demeaned and squared, and then the resulting value is detrended by subtracting a backward moving average for the

previous 60 days⁴⁴. The ‘Calendar’ variable denotes five dummy calendar variables, specifically a ‘January’ variable (set to 1 in the months of January and 0 otherwise), and day-of-the-week dummy variables for Monday (set to 1 on Mondays, 0 elsewhere), Tuesday, Wednesday and Thursday⁴⁵. Like Tetlock (2007) I also include a ‘Crash’ variable, which denotes the Black Monday flash crash of 19th October 1987. On this day the DJIA fell by 22.61%, which naturally represents a big outlier in the series. The crash variable therefore functions as a dummy variable, set to 1 on the 19th October 1987 and 0 on every other day. Finally, the ‘negative’ variable represents the standardized fraction of negative terms contained in the AofM column on day t . The standardization procedure was computed in the same manner as Tetlock (2007). First, the fraction of negative terms on day t is computed by dividing the number of negative terms on day t by the total number of terms on day t . This relative frequency is then demeaned by subtracting the day-of-the-week means of the prior year, and then standardized by dividing by the standard deviation of the day-of-the-week counts for the prior year⁴⁶. The purpose of this standardization is to ‘minimize semantic and stylistic noise’ such that the ‘conditional means are equal across different days of the week’ (Tetlock (2007)). As a result, the standardized sentiment variables become unitless, allowing us to interpret a change in negative sentiment as a change in standard deviation units. As in Tetlock, I also standardize the other continuous variables, by demeaning them and dividing by the standard deviation. This enables me to compare all variables with one another, as all continuous variables have been transformed to the same units. The regression equation is fitted using Ordinary Least Squares procedure, and Newey-West Robust Standard errors are used to account for heteroscedasticity and autocorrelation in the residuals up to 5 lags.

4.2.4 Model Estimation

In addition to validating and extending Tetlock’s results, this section also aims to explore different methods for dealing with polysemous entries in the GI dictionary. Existing studies (e.g. Tetlock (2007)) have either used the original GI system with its

⁴⁴Tetlock (2007) notes that other volatility proxies such as the VIX index give similar effects.

⁴⁵Friday is excluded to avoid issues of multicollinearity.

⁴⁶Thus for each Monday in 1985 the mean of all Mondays in 1984 is subtracted and the result is divided by the standard deviation of all Mondays in 1984. The result of this standardization means that the first year of observations are lost.

built-in disambiguation procedures, or else have not addressed the issue of polysemy at all. In the latter case it is often unclear as to how exactly the authors dealt with ambiguous terms and what effect - if any - this had on their sentiment proxy. Consequently, in this section various approaches to dealing with ambiguity are explored, and their impact related back to Tetlock’s study.

Table 4.1: Different senses for the word ‘capital’ in the GI dictionary, ordered from most-least frequent. The tags, POS information and definitions are taken directly from the corresponding entries in the GI dictionary. Here, only the semantic (content) tags are shown, but there are also non-semantic tags not shown here that are used as part of the GI’s disambiguation procedures. The tags Econ@ and Polit@ denote words of an economic or political orientation respectively; the tags WltOth and WltTot denote words to do with wealth; the tag Legal denotes words relating to legal, judicial or police matters; the tags PowAuth and PowTot denote words to do with power and the tag Ovrst represents words indicating overstatement. The other tags of Strong, Virtue, Negative and Vice are familiar from previous discussions, and the tag NgTV represents a version of the Negative category that was used in earlier versions of the GI system.

Sense	Tags	POS	Definition
1	STRONG, VIRTUE, ECON@, POLIT@, WLTOTH, WLTOT	Noun	Wealth, source of profit, moneyed interests; a center of government
2	NEGATIVE, NGTV, VICE, LEGAL, POWAUTH, POWTOT	Adj	Punishable by death. Idioms: ‘capital punishment’, ‘capital crime’, ‘capital charge’.
3	[None]	Noun	For emphasis. Idiom: ‘with a capital’
4	OVRST	Noun-Adj	Uppercase

As a running example consider the word ‘capital’, which has four different senses in the GI dictionary. These senses are listed in Table 4.1 in descending order from most-least frequent. The word ‘capital’ was chosen as a running example as it was one of the examples listed by Loughran & McDonald (2011, 2015) in their criticism of the GI dictionary. The authors claimed that the dictionary was unsuitable for analysis of financial texts, as negative words such as ‘capital’ were not typically negative in the domain of financial texts. However, in their criticism the authors failed to account for the fact that the dictionary comprises individual word senses. Table 4.1 shows that

only one of the four ‘capital’ senses is tagged as negative (in this case words occurring as part of the idiomatic expression ‘capital punishment’ or ‘capital crime’) whereas the other three senses are not. Distinguishing between senses purely on the basis of part-of-speech (POS) information is not always a viable option: three of the four senses for ‘capital’ are nouns, and therefore the correct sense cannot be resolved on the basis of POS information alone.

The question remains of how to distinguish the ‘correct’ sense from all of these available senses. Kelly & Stone (1975) (the authors of the GI’s disambiguation rules) conceive of four possible approaches:

- **Method 1: Ignore ambiguous terms completely.** Conduct the analysis using only unambiguous (monosemous) words, effectively removing ambiguous terms from the GI dictionary. Thus the word ‘capital’ would be ignored each time it is encountered in a given text.
- **Method 2: Treat ambiguous senses the same.** Collapse the different senses into a single sense, aggregating each of the assigned tags. Therefore each time the word ‘capital’ is encountered in text, apply all of the tags in Table 4.1 regardless of which sense is encountered. This would effectively treat all instances of ‘capital’ as if they were both negative and virtue.
- **Method 3: Apply the most frequent sense only.** The GI dictionary orders entries from most-least frequent. Thus each time the word ‘capital’ is encountered in text one could simply apply the tags associated with the most frequent sense (i.e. Sense 1). This would ensure that, on average, the correct sense would be applied.
- **Method 4: Use the GI disambiguation rules.** As previously discussed, the GI dictionary is designed to be used in conjunction with a set of disambiguation rules (Kelly & Stone (1975)). These rules examine the context around a particular key word to search for others words, phrases or semantic tags that could be used to disambiguate the word and apply the most likely sense.

These four methods for dealing with ambiguity are likely to result in different calculations of the sentiment variable. Specifically, the sentiment variable could be overestimated or underestimated depending on the method chosen. To explore what effect this

might have on the regression estimates, I calculate four regression models for Equation 4.1, one for each of the methods described above. The results of each of these regressions are presented in the ensuing section.

4.2.5 Experimental Results

4.2.5.1 The Impact of Negative Terms on DJIA Returns

The results of the OLS regressions for Equation 4.1 are shown in Table 4.2, along with the original results obtained by Tetlock. The table shows the estimated coefficients of the lagged negative terms against returns of the DJIA (the estimated coefficients of the other variables in Equation 4.1 are not shown here). The resulting coefficients are shown in basis points, where 1 basis point equals 1/100th of a percent. The results of a joint hypothesis test are also shown which tests for the joint contribution of the entire negative variable as a whole (all 5 lags) in explaining changes in returns of the DJIA. This is illustrated in the row marked $\chi^2(5)[Joint]$, with significant coefficients market in bold.

Method 1

In Method 1 ambiguous terms are ignored completely by removing ambiguous entries from the GI dictionary. This effectively reduces the number of negative entries in the dictionary from 2,291 to 1,711. While the estimated coefficients (column 1) produce the same signs as Tetlock’s results (column 5), none of the coefficients are statistically significant. The absolute magnitudes of the coefficients are also much lower. A joint hypothesis test for inclusion of the negative variable as a whole yields a statistically insignificant chi-square value of 7.7, which is much lower than the other models. This suggests that removing the ambiguous terms from the GI dictionary significantly underestimates the effect of negative sentiment on DJIA returns. This in fact is a logical conclusion as according to Crystal (2010) the number of meanings of a word is directly proportional to its frequency. Thus removing the ambiguous terms from the GI dictionary fails to account for the most frequently occurring sentiment terms.

Table 4.2: Regression of negative sentiment against DJIA returns. The table shows the estimated coefficients of the 5 lags of the negative variable in Equation 4.1. Counts of negative terms were generated using the system described in Chapter 3. Each Method 1-4 denotes one of the four methods for dealing with ambiguity. Method 1 involves ignoring ambiguous terms completely. Method 2 involves treating ambiguous senses the same. Method 3 involves applying the most frequent sense only. Method 4 involves using the GI’s disambiguation procedures. For each method a separate regression equation is computed using the frequency of terms computed using the different aforementioned disambiguation procedures. The coefficients in the table measure the impact of a one standard deviation change in negative sentiment on DJIA returns (results here are shown in basis points). The regression models are based on 3,709 observations from 01/01/1984 through 17/09/1999. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals up to five lags. Coefficients in bold are significant at the 5% level, and coefficients in bold and italics are significant at the 1% level.

Regressand: DJIA Returns					
Variable	Method 1	Method 2	Method 3	Method 4	Tetlock
<i>Negative_{t-1}</i>	-1.6	-3.6	-6.1	-4.4	-4.4
<i>Negative_{t-2}</i>	1.7	2.1	1.8	2.3	3.6
<i>Negative_{t-3}</i>	-1.1	-2.2	-2.1	-2.4	-2.4
<i>Negative_{t-4}</i>	2.5	3.4	4.7	3.8	4.4
<i>Negative_{t-5}</i>	1.9	2.8	2.6	2.9	2.9
$\chi^2(5)[Joint]$	7.7	16.1	28.2	19.8	20.8

Method 2

In Method 2 the GI’s polysemous senses are collapsed into one, effectively treating all of the senses the same and reducing the number of negative entries in the dictionary to 2,005. Consequently if one or more of the senses is tagged with a negative tag, I treat all word occurrences as if they are negative. This effectively overestimates the sentiment by maximizing recall at the expense of precision. The results obtained using this model are relatively similar to those obtained in Tetlock (2007), albeit with slightly smaller magnitudes and significance values. The chi-square statistic is also statistically significant, albeit slightly smaller than that of Tetlock’s. While Model 2 is a significant improvement over Model 1, it still slightly underestimates the sentiment relative to Tetlock. Furthermore, there is no guarantee that this model would produce similar results on a different data set.

Method 3

In Method 3 only the most frequent sense of a word is used, reducing the number of entries to only 1,892. As a result, the majority of the time the ‘correct’ sense is applied. The results suggest that all lags 1-5 have similar signs and significance values to those observed in Tetlock’s study. Additionally, the joint hypothesis test yields a highly statistically significant chi-square value of 28.2. Interestingly, all of the significant statistical tests (both the individual t-tests and the joint hypothesis test) all have much larger absolute magnitudes than Tetlock’s results. Conversely, all of the statistically insignificant statistical tests have lower absolute magnitudes. This result could therefore be potentially overestimating the effect of sentiment relative to the ‘true’ effect.

Method 4

In Model 4 the GI disambiguation rules from Kelly & Stone (1975) are used to attempt to resolve ambiguous words to their appropriate sense. This utilises all 2,291 negative entries in the GI dictionary, together with all associated disambiguation rules. Given that Tetlock used the original GI system (which encompassed both the GI dictionary and associated disambiguation rules) one would expect the results to be more-or-less identical. A comparison of the fourth column in the table reveals that the results are indeed very similar: lags 1, 3 and 5 of the sentiment variable have identical signs, magnitudes and significance values, and the chi-square value is much closer to Tetlock’s results than the other three models. There are some slight differences in the magnitudes of lags 2 and 4, and the chi-square value is slightly lower than in Tetlock (2007), but these small differences are likely due to differences in data pre-processing, standardization of variables, rounding errors, or the order in which the calculations are produced. Despite attempts to contact the author about alignment issues⁴⁷ and standardization issues, he was unable to provide specific details regarding his implementation, and so these issues were left unresolved.

⁴⁷Specifically, I attempted to contact the author about his procedure for dealing with days where there was news but no returns, and vice versa.

Discussion

Overall, it is evident that the system is capable of re-producing Tetlock’s results. While Methods 1-3 may suffice for certain applications, the closest statistical comparison is found using Method 4, where the GI’s disambiguation procedures are used to identify the most appropriate sense. As a result, the disambiguation routines (Method 4) will be used throughout the remainder of this thesis.

The fact the estimated coefficients at lags 1, 3 and 5 in Method 4 are identical to those reported in Tetlock (2007) is extremely encouraging. Although there are some minor differences in the magnitudes of the coefficients at lags 2 and 4, the signs and significance of the coefficients at these lags are the same. There is a minor difference in the magnitude of the joint hypothesis test, with Tetlock reporting a chi-square value of 20.8 vs. the value of 19.8 in this thesis. However, given that the chi-square value is equally significant and so close in proximity does not lead to cause for concern. Despite the fact that the system used here is built independently of the system used by Tetlock (2007), it has been possible to reproduce quantitatively analogous results. Not only does this validate the construction of the system, and also provide independent confirmation of Tetlock’s results, but it also validates understanding of the method used. Consequently, in the remainder of this chapter, where the impact of the vice terms is explored, one can rest assured that any significant differences are systematic and not due to measurement or model errors.

In the interest of being fastidious, I now consider some potential causes for the slight discrepancies between Tetlock’s results and the closest-matching results using Method 4. Firstly, it is not clear how Tetlock pre-processes his texts. Here, a custom-built program was used to strip the document of any unwanted meta-data, but the exact meta data removed may have differed from that removed in Tetlock’s study. Secondly, words and sentences in this system are tokenized using the Stanford Core NLP Toolkit (Manning et al. (2014)). It is not documented how the GI system tokenized sentences or words, and so there are likely to be some minor discrepancies between total word counts. Thirdly, it is not clear how Tetlock aligned the disparate dates in the different time series. There were a number of days when news was written, but no financial information was available (such as public holidays), and conversely a number of days when there was financial information but no news. Attempts to contact the author unfortunately did not elucidate this issue, and so in this thesis

these disparate observations were simply removed. This resulted in a slightly shorter number of observations than used by Tetlock. Finally, the order of computing the statistical calculations naturally affects the estimated coefficients. For instance, should the counts of negative words be standardized before or after dividing by the total number of words in the text? And should the standardization procedure be conducted after removing disparate observations? Again, attempts to contact the author did not shed light upon this issue. Collectively, these issues may have accounted for some of the differences between figures, or else the differences may have been attributed to some other unknown factor.

4.2.5.2 The Impact of Vice Terms on DJIA Returns

Having replicated the results in Tetlock’s study I now turn towards extending his work. The main premise of this thesis is to test whether negative sentiment is rooted in the ethical value system. It has already been established that the vice terms in the GI dictionary constitute a direct subset of the negative terms. Consequently, I now examine whether it is this subset of vice terms that are really responsible for changes in returns. To do so, the VAR equation in the preceding section (namely Equation 4.1) is recomputed using the GI’s vice terms as a proxy for investor sentiment. The data and methods used here are identical to those in Section 4.2.5.1. However, rather than using counts of the GI’s negative terms, counts of vice terms are used instead. These counts are standardized in the same manner as the negative terms (described in the previous section).

Table 4.3 shows descriptive statistics for the DJIA returns, and the fraction of negative and vice terms in the AofM corpus, which are all normalized to percentage scores to enable comparison. All statistics shown in the table are calculated using the variables’ pre-standardized values. Note that the Black Monday crash of 19 October 1987 was removed from the descriptive statistics. On this day the DJIA fell by 22.61% and consequently this observation represents a significant outlier which distorts the true kurtosis and skewness values⁴⁸.

The results show that the mean and median values for the negative variable (2.62%

⁴⁸Kurtosis measures the fatness of a distribution’s tails and its peakedness about the mean (Brooks (2014)). The skewness defines the shape of a given distribution, measuring its symmetry about the mean (Brooks (2014)).

and 2.53% respectively) are much higher than the equivalent values for the vice variable (0.48% and 0.45%). This is perhaps unsurprising given that the vice terms comprise a subset of the negative terms. The standard deviation is also much higher for the negative terms than the vice terms (0.84% vs. 0.28% respectively), suggesting that the negative terms exhibit almost three times more variability than the vice terms. In addition the vice terms are more leptokurtic than the negative terms, exhibiting a kurtosis value closer to that of returns than to the negative terms. The skewness of the vice terms is also much higher than the skewness of the negative terms. In terms of the returns, the mean and median values are approximately 0. However the high, positive kurtosis value of the returns suggests the presence of extreme returns is likely, even after removal of the Black Monday anomaly.

Table 4.3: Descriptive statistics in the ‘Abreast of the Market’ column (WSJ) during the period 01/01/1984 through 17/09/1999. Returns denotes the continuously compounded returns of the DJIA (the Black Monday crash of 19/10/1987 was identified as a significant outlier and consequently was removed from these calculations). The negative and vice variables denote the fraction of negative/vice terms contained in the column. All three variables are normalized to percentage scores. The total number of observations is 3,963.

Variable	Mean	Median	Std. Dev.	Kurtosis	Skewness	Min	Max
Returns	0.00	0.00	0.01	3.44	-0.44	-0.08	0.05
Negative	2.62	2.53	0.84	0.15	0.49	0.55	6.17
Vice	0.48	0.45	0.28	9.36	1.46	0.00	4.00

Given that the vice terms constitute a subset of the negative terms one might expect them to exhibit similar effects on returns. Consequently I recompute the VAR models in Equation 4.1, substituting counts of negative terms with counts of vice terms. Table 4.4 shows the results of this substitution, alongside the full set of negative terms for comparison.

The results show that the signs of the coefficients of the vice variable at lags 1, 3, 4 and 5 are the same as the signs of the equivalent lags for the negative variable. However, the vice variable at lag 2 has a different sign from the equivalent lag 2 for the negative variable. The magnitudes of the coefficients of the vice variable are also much smaller than the equivalent magnitudes of the negative variable, and indeed none of the lags are statistically significant at the conventional levels. The joint hypothesis test

is also statistically insignificant, resulting in a low chi-square value of 4.5, relative to the value of 19.8 reported for the negative variable in Table 4.2. Overall, these results suggest that the impact of vice terms in general financial news is not as strong as the impact of negative terms, and that in this data set the vice terms cannot simply be substituted for the negative terms.

Table 4.4: Regression of vice terms against DJIA returns. This table shows the estimated coefficients of the 5 lags of the vice variable (which substitutes for the negative variable in Equation 4.1), along with the equivalent negative results from Table 4.2 for reference. The vice column in the table denotes the standardized fraction of vice terms in the AofM corpus, and the negative column denotes the standardized fraction of negative terms. The coefficients measure the impact of a one standard deviation change in vice terms on DJIA returns. The regression model is based on 3,709 observations from 01/01/1984 through 17/09/1999. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals upto five lags. Coefficients in bold are significant at the 5% level, and coefficients in bold and italics are significant at the 1% level.

Regressand: DJIA		
Lag	Negative	Vice
<i>Vice_{t-1}</i>	-4.4	-1.6
<i>Vice_{t-2}</i>	2.3	-0.3
<i>Vice_{t-3}</i>	-2.4	-2.3
<i>Vice_{t-4}</i>	3.8	1.1
<i>Vice_{t-5}</i>	2.9	0.5
$\chi^2(5)[Joint]$	19.8	4.5

4.2.5.3 Exploring the Time-Varying Impact of Sentiment

So far this chapter has shown how the system described in Chapter 3 can be used to reproduce the results of a major existing study. Using the data and methods from Tetlock (2007) it was shown how negative sentiment expressed in the WSJ’s ‘Abreast of the Market’ opinion column has a significant impact on next day returns at the index level. The study was subsequently extended to the GI’s vice terms, however it transpired that these terms did not have a significant impact on next-day DJIA returns. In this section the investigation is extended further, by exploring the time-varying impact of the sentiment proxy. Specifically, a series of rolling regression estimations are computed to see whether the impact of negative and vice terms is confined to

specific periods of time.

The motivation behind this section is that the coefficients estimated in the previous section are unlikely to remain constant over time. In reality financial markets are dynamic in nature and evolve from one period to the next. Performing rolling regressions takes account of this dynamic nature, by allowing the parameters of the model to vary over time. It may be the case that there are only certain periods in which the negative and vice variables are able to explain some of the variability in market returns. Thus instead of assuming that the market is (or is not) always efficient, this method allows for the fact that market may be inefficient at certain times.

The data used in these regression equations is the same of that in the preceding sections. Specifically, the AofM column from the Wall Street Journal is used for the period 1984-1999. To test the time-varying impact of sentiment a rolling regression approach is adopted using a window size of 130 days (corresponding to approximately 6 months of trading days). This rolling window is moved forward one day a time, such that each rolling window has an overlap of 129 observations with the preceding window. In each rolling window the fitted model for Equation 4.1 is computed, both for the negative terms and the vice terms. This resulted in two sets of 3,558 rolling regression models: one in which the negative variable is used and one in which the vice variable is used. If market sentiment was really being driven by ethical evaluations, then one would expect the periods in which negative and vice terms are able to account for variability in returns to be broadly similar.

In each rolling window a t-test is conducted on the leading (first) lag of the negative or vice variables. The reason for testing the first lag rather than a joint hypothesis test is because in Table 4.2 it was the first lag of negative sentiment that was found to have a significant impact on next day returns. Therefore, in each rolling window a t-test is computed for $Sentiment_{t-1}$, and the associated p -value is recorded. This results in two sets of p -values (one for the vice variable and one for the negative variable). After computing the requisite models, the p -value associated with each coefficient is extracted, and subsequently binarised with the following surjective function:

$$f(x) = \begin{cases} 1, & \text{if } p \leq 0.05 \\ 0, & \text{otherwise} \end{cases} \quad (4.2)$$

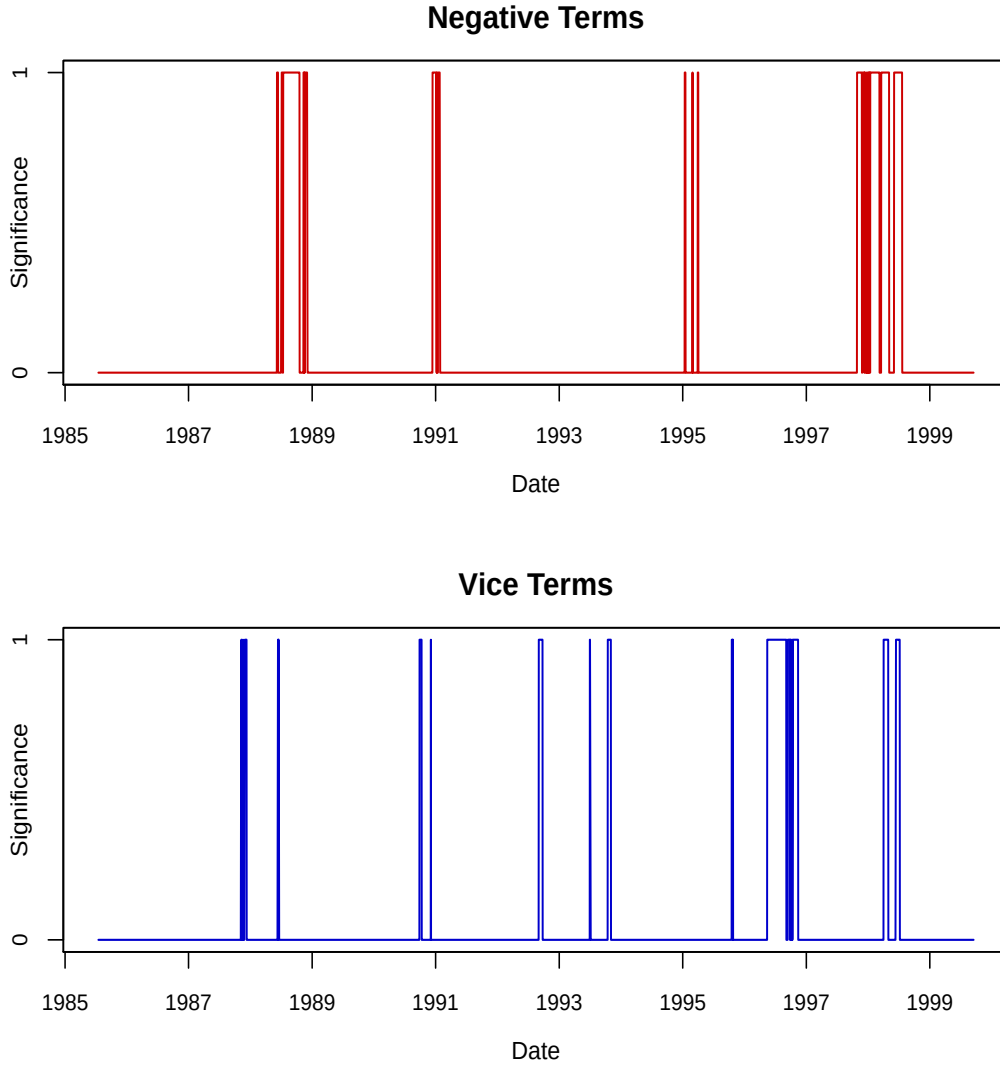
In other words, p -values that are statistically significant at $p < 0.05$ are given a value of 1 and p -values that do not meet this criterion are given a value of 0. This allows for easy visualisation of the results.

The results of the rolling regression study are shown in Figure 4.2. The first notable observation is that the impact of negative and vice terms on DJIA returns is episodic. That is, there are certain periods where the negative/vice terms have a statistically significant impact on next-day returns, and certain periods where they do not. Out of the 3,558 rolling models computed, the first lag of the negative variable ($negative_{t-1}$) is statistically significant in 259 models (7.28%). Conversely, the first lag of the vice variable ($vice_{t-1}$) is statistically significant in 211 models (5.93%). This would seem to suggest that the negative variable contains additional information about market-level returns that is not found using only the vice terms.

The second notable observation is that there is no overlap between the two models. Despite the fact that the vice terms in the GI dictionary comprise a subset of the negative terms, the periods in which the vice terms are statistically significant do not overlap with the periods in which the negative terms are significant. This suggests that vice terms cannot simply be substituted for negative terms, and that they are not responsible for the observed effect between negative sentiment and financial returns. It also shows that the vice terms contain useful information about returns that is lost when using the broader set of negative terms.

Another key point to note is that while the vice terms were not significant across the entire time period (1984-1999) there are certain time periods when the vice terms are significant. Conversely, while negative terms were globally significant across the entire period, in the rolling regression models there were certain periods where the negative terms were not significant. This suggests that market efficiency is episodic, in that there are certain periods where negative/vice terms can account for changes in price that cannot be accounted for using other market variables. During these periods, returns were dependent on evaluative or ethical measures, suggesting that markets were driven by investor sentiment rather than underlying financial variables.

Figure 4.2: Rolling regression estimations using negative and vice terms from the GI dictionary. The rolling regressions (N=3,558) were conducted using data from the ‘Abreast of the Market’ column in the Wall Street Journal (01/01/1984 - 17/09/1999). The rolling regression uses a 130-day rolling window, moving the window forward by one day at a time. A t-test is conducted on the first lag of the negative or vice variable in each period. The figures here show the significant estimated regression models, with the red line indicating negative terms and blue line indicating vice terms. A significance of 1 means $p < 0.05$ for the particular rolling window. A significance of 0 means $p \geq 0.05$.



4.2.6 Summary of the Market-Level Study

In this initial case study the system was validated by replicating the results of a major existing study. Specifically, the study replicated the results of Tetlock (2007), who explored the impact of negative sentiment in the WSJ’s ‘Abreast of the Market’ column. Four separate models were tested, demonstrating various approaches to dealing with ambiguity in the GI dictionary. While there were some differences in the system used

in this thesis and the one used in Tetlock (2007), it was shown that the system was able to reproduce the results from the previous study. Overall, it was found that a unit increase in negative terms led to a -4.4 decrease in DJIA returns the following day.

After successfully replicating the results of this previous study the experiment was then extended to the ethical vice terms. The same methods, data and time period were used, except in this experiment it was now the frequency of vice terms being modelled. The results of this study suggested that in the overall period (01/01/1984 through 17/09/1999) the vice terms had no statistically significant impact on DJIA returns, either at any of the individual lags (1-5) or at the joint level. This shows that the vice terms cannot be simply substituted for the negative terms.

Following this, a series of rolling regressions were then computed, in order to explore the time-varying impact of the two sentiment variables. It was discovered that the impact of both negative and vice terms on next-day DJIA returns was episodic, suggesting that the terms were able to occasionally explain some of the variability in returns that could not be accounted for by any other market variable. Interestingly, there was no overlap between the periods in which the vice terms were significant and the period in which the negative terms were significant, suggesting that each variable was measuring a separate construct. Each variable therefore contained useful information about returns that was lost when using only one of the variables in isolation.

4.3 Case Study 2: Sentiment at the Firm-Level

4.3.1 Introduction

In this section I build upon the work of the previous case study, by exploring the effect of negative and vice terms on firm-level returns. The former Enron scandal is used as a case study on the basis that the company and its executives were engaged in a wide variety of criminal practices, and thus the news is likely to be laden with ethical terminology. The work in this section expands upon work that was published in Cook & Ahmad (2015). In Cook & Ahmad (2015) initial evidence was provided of the relationship between ethics and affect, using negative and vice terms from the GI dictionary. Using a corpus of texts about the Enron Corporation a significant correlation was observed between the frequency distributions of negative and vice terms. A series of autoregressive models were subsequently computed in which returns were

regressed against their lagged values, and against the frequency of negative and vice terms on the preceding day. The results suggested that neither negative nor vice terms had any significant impact on next day returns. However, the initial study in Cook & Ahmad (2015) did not attempt to account for other sources of market variability, and did not include variables to account for market liquidity, market volatility or calendar effects as per Equation 4.3. Furthermore, the issue of the GI's disambiguation was ignored entirely, and a simple approach was taken whereby the ambiguous entries in the GI dictionary were collapsed onto a single sense (as per Method 2 in the preceding section). The implication of this decision was that all words would be tagged as negative or vice if any one of their senses in the GI dictionary was tagged as such. In this section I address these issues in addition to expanding upon this previously published work. In the remainder of this preamble a brief overview of the history of Enron is provided in order to motivate its inclusion in this study.

Enron: An overview

Enron was originally formed in Houston, Texas, following the merger of Houston Natural Gas and InterNorth in 1985. The company initially focussed on providing natural gas, before branching into other energy markets, the communications industry and the services industry. The company played a fundamental role in the deregulation of the electrical power markets in 1988 (Sims & Brinkmann (2003)), from which they prospered enormously. Throughout the 1990s the company exhibited impressive growth, with shares reaching an all-time high of \$90 per share in August 2000. In addition to strong economic performance, the company was also widely admired in both public and governmental spheres. For six consecutive years the company was labelled as '*America's most Innovative Company*' (Fortune Magazine), and was heralded for its philanthropic efforts and honouring of leading world figures such as Nelson Mandela.

However, in 2001 things turned sour and Enron's 'house of cards' came tumbling down. By November 2001 the company's share price had plummeted to less than \$1 per share, and in December 2001 the company was forced to file for bankruptcy protection - the largest in US history at that time. A subsequent investigation by the US Securities and Exchange Commission (SEC) found that key company executives had been creating a number of special purpose entities to manipulate financial reporting by inflating earnings and hiding losses from the books (Baker et al. (1993)). A number of

indictments of several senior executives ensued, and 20,000 employees lost their jobs. Many employees also lost their life savings, having been encouraged to buy into Enron stock by the former management. Enron existed as a company for 16 years, and was once heralded for its virtuous business practices. However, after its fall the company was only remembered by morally negative terms - the vice words. The transition from a virtuous company to a vicious one did not happen overnight, and it is possible that there were traces of Enron's illicit business practices in the news before the share price plummeted. The question is whether this criticism involved the use of vice terms in contrast to the positive appreciation previously expressed.

4.3.2 Data

The textual data used in this study comprises six newspaper and newswire sources obtained through the LexisNexis and ProQuest databases. The sources were chosen partly based on circulation data (Circulation data was obtained from the Alliance for Audited Media (AAM), www.auditedmedia.com), partly on accessibility (not all of the AAM sources were available through LexisNexis/ProQuest) and partly on their coverage of Enron (to be included, each source had to have at least 50 articles about Enron during the covered time period). The selection criteria involved at least one mention of the names 'Enron', 'InterNorth', 'Houston Natural Gas' or 'HNG' in the title of the article, plus at least 3 additional mentions of either keyword in the main body of the text. The search was restricted to articles timestamped between January 1997 (the earliest period for which historical share price information was available) and December 2001 (the month that Enron filed for bankruptcy). In total this resulted in 1,776 articles after removal of exact duplicates⁴⁹. A breakdown of the number of articles from each source is given in Table 4.5.

Historical share price and volume information was obtained from Gilardi & Co's Legal Settlement website (Available from: <http://www.gilardi.com/pdf/enro13ptable.pdf>, Last Accessed 02 November 2016). Volume information was used as a liquidity proxy akin to the volume of the NYSE used in the previous section. The volatility proxy used in this section was the same as that of the previous case study: specifically, the

⁴⁹ Articles published on weekends were dropped from the corpus, and the frequency of negative and vice terms on days with no news was set to 0 to facilitate alignment of the news and financial time series data.

returns were demeaned to obtain a residual which was then squared and detrended by subtracting a backward moving average for the previous 60 days. Similar results (not reported here) are obtained using the VIX volatility index as a proxy for market risk. A time series plot of Enron's standardized returns is shown in Figure 4.3. These returns are standardized to z-values by subtracting the mean and dividing by the standard deviation. The plot shows a number of extreme observations are evident, such as on 16/08/1999 ($z = -9.24$), 28/11/2001 ($z = -25.55$), 29/11/2001 ($z = -7.03$) and 04/12/2001 ($z = 10.47$) amongst others. Results are presented in the ensuing section both with and without these outliers.

Table 4.5: Composition of the Enron corpus (01/01/1997 through 31/12/2001). Each text contains the word 'Enron', 'InterNorth', 'Houston Natural Gas' or 'HNG' at least once in the title, and at least 3 additional mentions in the body of the text. The Wall Street Journal and Financial Times articles were downloaded from the ProQuest news database and all other sources were downloaded from LexisNexis. The first column denotes the name of the source, the second the number of articles. The third column denotes the number of tokens in each source, and the fourth and fifth columns denote the number of negative and vice terms respectively.

Source	Articles	Days	Tokens	Negative	Vice
Associated Press	372	122	210,544	6,455	2,061
Business Wire	198	128	101,687	1,713	423
Financial Times	326	155	151,241	4,892	1,266
PR Newswire	525	346	308,507	3,554	745
New York Times	101	68	84,652	2,452	700
Wall Street Journal	254	185	173,774	4,444	1,172
Total	1,776	1,004	1,030,405	23,510	6,367
Mean	296	167	171,734	3,918	1,061

One of the main assumptions of regression using time series data is that the time series used are stationary. Table 4.6 shows the results of stationarity tests on each of the continuous time series variables used in the regressions. These tests used the Augmented Dickey-Fuller (ADF) test (described in the previous chapter) to test the null hypothesis of non-stationary data. As can be seen from this table, all ADF statistics yield a p -value of < 0.05 , rejecting the null hypothesis of non-stationarity.

Figure 4.3: Enron's returns for the period 01/01/1997 through 31/12/2001 (N=1,191). Returns have been standardized to z-scores to have a mean of 0 and variance of 1.

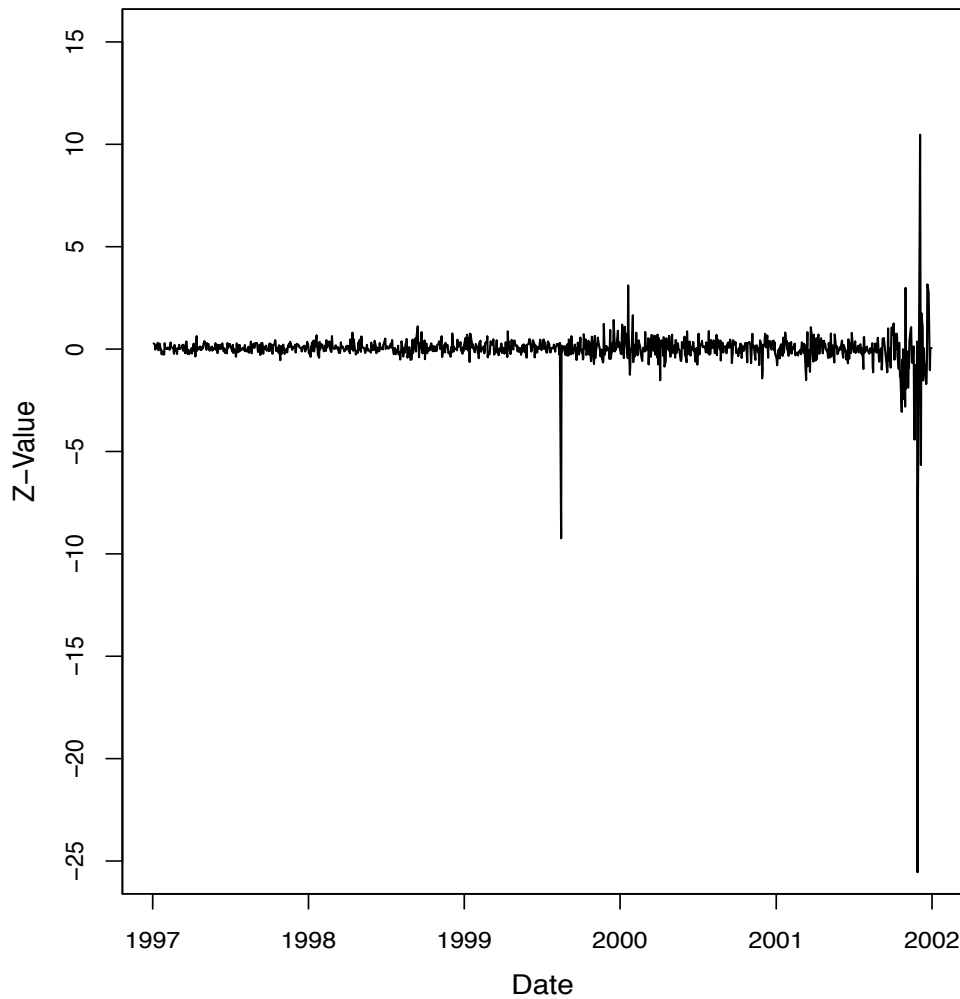


Table 4.6: Augmented Dickey-Fuller (ADF) tests for stationarity of continuous time series variables. The volume and volatility variables are log transformed and detrended. Stationarity is tested for variables up to five lags. The ADF statistic tests the null hypothesis of non-stationary time series, thus p -values of < 0.05 are observed to be stationary.

Variable	Dickey-Fuller Statistic	P-Value
Returns	-12.48	< 0.01
Volume	-6.15	< 0.01
Volatility	-11.19	< 0.01
Negative	-4.96	< 0.01
Vice	-4.89	< 0.05

4.3.3 Methods

Having discussed the composition of the dataset (both textual and financial) I now turn to a discussion of the modelling process. In this section the following VAR equation is fitted using OLS regression:

$$r_t = \beta_0 + \beta_1 L_5 r_t + \beta_2 L_5 Sentiment_t + \beta_3 L_5 Volume_t + \beta_4 L_5 Volatility_t + \beta_5 Calendar_t + \varepsilon_t \quad (4.3)$$

Where:

r_t = Enron's (winsorized) returns at time t

β_0 = a constant representing the mean returns when all other variables are set to 0

$Sentiment_t$ = denotes the frequency of the sentiment variable (negative or vice terms)

$Volume$ = the volume of Enron shares traded (log transformed and detrended)

$Volatility$ = a proxy for market risk (calculated using Enron's returns)

$Calendar$ = the January and day-of-the-week dummy variables

ε_t = a random error term with mean of 0

All continuous variables in the equation are standardized to have a mean of zero and unit variance. This standardization allows for direct comparison of the changes of different variables, and to interpret the results as changes in standard deviation units. The volume and volatility variables are log transformed and detrended in the manner previously described in the preceding section. Specifically, the volume variable is log transformed to stabilize the variance, and detrended by subtracting a backwards 60 day moving average. The volatility variability was constructed by demeaning Enron's returns, squaring the residual, and then detrended by subtracting a backwards 60 day moving average. Both approaches are robust to using different moving average windows (e.g. 30 days instead of 60 days).

It is prudent to acknowledge the existence of outliers in the Enron data set. Following an initial investigation by the SEC, and subsequent bankruptcy proceedings, Enron experienced a large decline in their share price. Figure 4.3 shows that there are a number of extreme values in the returns series, which could have a significant impact upon

the model's fit and standardization procedures. The largest of these extreme values occurred on the 28th November 2001 where Enron's share price reduced from \$4.11 per share to only \$0.61 per share - an 85% decrease. To reduce the impact of outliers the returns data are winsorized⁵⁰ in the upper and lower 1% tails of the distributions, replacing the extreme values with the 1st and 99th percentiles. Results are subsequently presented both with the winsorized returns and without them to demonstrate whether these extreme observations introduce any statistical bias.

4.3.4 Experimental Results

4.3.4.1 Descriptive Statistics

Descriptive statistics for the frequency of negative and vice terms and for Enron's returns (both winsorized and non-winsorized) are shown in Table 4.7. The table shows that the frequency of negative terms in the Enron corpus is approximately 5 times higher than the frequency of vice terms, as evidenced by the means of 1.55% and 0.31% respectively. The fact that the mean for both the negative and vice variables is slightly higher than the median values, suggests also that the underlying frequency distributions are positively skewed. This fact is supported by the skewness statistics which are positive for both variables. This skewness is perhaps unsurprising given that these variables represent relative frequency data, which has a fixed lower (and upper) bound. The standard deviation of negative terms is also higher than the standard deviation of vice terms, suggesting that the negative terms exhibit more variability than the vice terms. While the distribution of both negative and vice variables is leptokurtic (relative to the normal distribution) there is slightly more kurtosis in the vice variable than the negative one. Overall, the descriptive statistics in Table 4.7 led to similar conclusions to those drawn in Table 4.3 in the previous case study.

In terms of the returns, the average return over the period 01/01/2007 to 31/12/2001 is -0.34%, compared to the value of zero that would be expected in an efficient market. The returns distribution is also exhibiting strong signs of negative skewness, as evidenced by the skewness statistic of -13.41 and by the fact that the mean value is lower

⁵⁰Winsorization is a form of outlier treatment used to 'robustify statistical inferences' (Ruppert (2004)). Rather than removing extreme values, the winsorization process sets their values to a pre-determined percentile of the data sample. Here I use a 98% winsorization, setting values below the 1st percentile and above the 99th percentile to their 1st and 99th percentile values respectively.

than that of the median value. These statistics support the notion that the returns series contains a large number of potential outliers as evidenced by the high kurtosis value and extreme minimum and maximum values. These outliers were also evident visually in Figure 4.3.

Table 4.7: Descriptive statistics for the Enron corpus during the period 01/01/1997 through 31/12/2001. Returns denote Enron’s continuously compounded returns. Wins. returns denote trimmed returns at the upper and lower 1% of the returns distribution. The negative and vice variables denote the fraction of negative/vice terms contained in the corpus. All variables are multiplied by 10^2 for ease of comparison.

Variable	Mean	Median	Std. Dev	Skewness	Kurtosis	Min	Max
Returns	-0.34	0.00	7.45	-13.41	356.74	-190.77	77.70
Returns (wins.)	-0.18	0.00	3.06	-1.07	4.75	-13.70	8.24
Negative	1.55	1.16	1.30	1.43	4.43	0	10.91
Vice	0.31	0.16	0.41	2.25	9.31	0	3.64

In this case study results are presented both with and without removal of outliers. Consequently, in addition to the ordinary returns, Table 4.7 also presents winsorized returns whereby the upper and lower tails of the returns distribution are trimmed at the 1% level. This effectively limits the influence of the outlier values, bringing the winsorized returns more in line with the results obtained in Table 4.3. From Table 4.7 it can be seen that the winsorization process significantly reduces the skewness, kurtosis and standard deviation values, and also has some effect on the mean.

4.3.4.2 The Impact of Negative and Vice Terms on Enron’s Returns

Winsorized Returns

It was previously noted that Enron’s returns series contained several outlier values. These outliers could alter the overall mean of the returns series and disproportionately affect the hyperplane that is fitted by the OLS regression procedure. This in turn could result in an underestimation or overestimation of the ‘true’ coefficients of the regression model. To avoid these outliers from unduly affecting the regression estimates the upper and lower tails of the returns distribution are first winsorized. Specifically, the lower and upper 1% of the distribution are trimmed, such that values less than the 1st percentile value are replaced with the 1st percentile value and values greater than the 99th percentile value are replaced with the 99th percentile value. This effectively caps

the influence of the lower and upper 1% of observations. Table 4.8 shows the estimated coefficients of the *Sentiment* variable in Equation 4.3 using these winsorized returns as the dependent variable. The table shows the estimates of two regression models: one in which the *Sentiment* variable denotes the (standardized) counts of negative variables, and one in which the variable denotes the (standardized) counts of vice variables. The counts were constructed using the system described in Chapter 3, without using any external override dictionary. The GI dictionary and disambiguation rules were used to resolve words to their correct senses as per Method 4 in the previous case study. In both cases the coefficients were estimated using OLS regression, with Newey-West robust standard errors to account for heteroscedasticity and autocorrelation in the residuals.

Table 4.8: Regression of negative and vice terms against Enron winsorized returns during the period 01/01/1997 through 31/12/2001 (N = 1,191 days). The table shows the estimated coefficients (in basis points) of two models: one in which the sentiment variable in Equation 4.3 is denoted by the count of negative variables and one in which it is denoted by the count of vice variables. In both cases the counts were standardized to z-values by subtracting the mean and dividing by the standard deviation. The estimated coefficients in the table are shown in basis points, and Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals. A dot after the coefficient represents significance at the 10% level. A bold coefficient denotes significance at the 5% level. Bold and italics denotes significance at the 1% level.

	Negative	Vice
Lag_{t-1}	-10.1	<i>-13.4</i>
Lag_{t-2}	-7.2	-4.7
Lag_{t-3}	-2.4	-4.1
Lag_{t-4}	-10.6	-8.8.
Lag_{t-5}	-5.8	-2.6
$\chi^2(5)[Joint]$	<i>16.44</i>	<i>14.99</i>
MSE	0.896	0.897

For the negative model, denoting the standardized frequency of negative terms, negative coefficients are observed for all lags 1-5. This suggests that an increase in negative terms is followed by a subsequent decrease in Enron's returns. The fact that this effect is observed at all lags 1-5 implies that the impact of negative terms upon returns is non-transient, in that it is not reversed over subsequent days. This contrasts

with the results in the market-level case study where the initial impact of negative terms was followed by a subsequent reversal. Of particular note are the coefficient estimates at lags 1 and 4, which are -10.1 and -10.6 respectively and are both statistically significant at $p < 0.05$. These coefficients imply that a unit increase in negative terms on days $t - 1$ or $t - 4$ lead to a -10.1 or -10.6 basis point decrease respectively in returns on day t . A joint hypothesis test for the inclusion of all lags of the negative variable is also statistically significant ($\chi^2 = 16.44, p < 0.01$), suggesting that the five lags of the negative variable jointly impact upon Enron's returns.

The vice model yields similar conclusions to the negative regression model: all lags 1-5 are negative (suggesting that an increase in vice terms is followed by a subsequent decrease in Enron's returns), and this effect is non-transient. The estimated coefficient of the vice variable at lag 1 is highly statistically significant ($p < 0.01$), and suggests that a unit increase in the prior day's vice terms results in a -13.4 basis point decrease in next day returns. This effect is slightly stronger than the negative model, implying that some information might be lost when using the broader set of negative terms. While lag 4 of the negative model was statistically significant (at $p < 0.05$) this is not the case for the vice variable, which is only marginally significant (at $p < 0.10$). The joint hypothesis test of the vice model is also statistically significant, suggesting that all lags 1-5 of the vice variable jointly impact upon the returns. The effect of negative and vice variables here is very similar, suggesting that vice terms may be substituted for the negative terms without too much loss of predictability. Indeed this is supported by the similar MSEs, suggesting that the two models offer a comparable fit.

Non-Winsorized Returns

Having established that negative and vice terms impact upon Enron's winsorized returns, I report now on the equivalent results using the original non-winsorized returns. Table 4.9 reports on the estimation of Equation 4.3 using the original non-winsorized returns as the dependent variable. As before, two models are estimated - one denoting the standardized frequency of negative terms and one denoting the standardized frequency of vice terms. All other model parameters remain as previously described. The results show that the winsorization process clearly had a significant effect upon the estimated regression coefficients. In Table 4.9, where the dependent variable denotes Enron's non-winsorized returns, neither the negative nor the vice model are jointly sta-

tistically significant, whereas both models were jointly significant in Table 4.8. While the estimated coefficients were negative for all lags 1-5 in the winsorized models, the results using the non-winsorized returns result in a positive coefficient at lag 3 for both the negative and vice models. Furthermore, while the negative model in Table 4.8 yielded statistically significant coefficients at lags 1 and 4, and the vice model yielded a statistically significant coefficient at lag 1, none of the equivalent lags in Table 4.9 are statistically significant. Instead, the only statistically significant coefficient in Table 4.9 is lag 2 of the negative variable.

Table 4.9: Regression of negative and vice terms against Enron's non-winsorized returns during the period 01/01/1997 through 31/12/2001 ($N = 1,191$ days). The table below shows the estimated coefficients (in basis points) of two models: one in which the sentiment variable in Equation 4.3 is denoted by the count of negative variables and one in which it is denoted by the count of vice variables. In both cases the counts were standardized to z-values by subtracting the mean and dividing by the standard deviation. The estimated coefficients in the table are shown in basis points. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals. A dot after the coefficient represents significance at the 10% level. A bold font denotes significance at the 5% level.

	Negative	Vice
Lag_{t-1}	-4.9	-3.0
Lag_{t-2}	-9.4	-5.8
Lag_{t-3}	3.7	2.0
Lag_{t-4}	-7.4.	-6.5
Lag_{t-5}	-4.8	-0.5
$\chi^2(5)[Joint]$	10.87	6.53
MSE	0.794	0.799

Overall these results show that the observed outliers had a large impact upon the estimated regression coefficients. Capping the influence of these outliers by winsorizing the upper and lower 1% of the returns distribution resulted in different coefficient magnitudes, signs and significance levels, suggesting that the impact of negative and vice terms on Enron's returns are sensitive to extreme observations and that the negative and vice terms had greater predictability of returns in the absence of such outliers. This may in part be due to the higher degree of volatility in the returns series: the standard deviation of Enron's returns is 0.07 relative to the standard deviation of 0.01 in the

DJIA returns of the first case study. However, it could also be attributable to the fact that the returns series is over a shorter time horizon, with only 1,191 observations vs. 3,963 observations in the market-wide case study.

4.3.4.3 The Impact of Returns on Sentiment

Having established that negative sentiment and ethical terms do impact firm-level returns, I turn now to address the issue of whether this relationship is bi-directional. It may be the case that the news reports upon events that have already been compounded into the share price, in which case changes in returns might contain useful information about the changes in sentiment. Alternatively, it might be the case that there is some sort of feedback phenomenon, whereby negative news causes a change in share price, which leads to more negative news, and so on. Consequently, in this section I posit the following equation:

$$\begin{aligned} Sentiment_t = & \beta_0 + \beta_1 L_5 r_t + \beta_2 L_5 Sentiment_t + \beta_3 L_5 Volume_t \\ & + \beta_4 L_5 Volatility_t + \beta_5 Calendar_t + \varepsilon_t \end{aligned} \quad (4.4)$$

This is equivalent to the model in Equation 4.3, except for the fact that the sentiment variable has replaced the returns variable as the dependent variable. The methods and data used are as in the preceding sections. The returns used are the winsorized returns, but similar conclusions are reached using the non-winsorized returns. The results of the two regression equations (one for the vice terms and one for the negative terms) are shown in Table 4.10.

In the negative model lags 2 and 3 are statistically significant at $p < 0.05$, yielding coefficients of -4.0 and -4.3 respectively. This implies that there is a bi-directional relationship between negative terms and returns, and that a unit increase in returns at time $t - 2$ or $t - 3$ yields -4.0 and -4.3 basis point decreases in the fraction of negative words at time t . Interestingly, while lags 2 and 3 are statistically significant in Table 4.10, it was lags 1 and 4 that were statistically significant in Table 4.8 implying that there some feedback phenomenon appears to be present: sentiment initially impacts returns, which in turn impact sentiment and which subsequently impacts returns. The chi-square statistic, which tests whether returns jointly impact upon the frequency of

negative terms, is also statistically significant at the 1% level, implying that returns collectively impact upon the frequency of negative terms. Finally the MSE of 0.382 is much smaller than the equivalent MSE in Table 4.8.

In terms of the vice model, only lag 3 is statistically significant at $p < 0.05$, implying that a unit increase in returns at time $t - 3$ leads to a -5.5 basis point decrease in terms on day t . The estimated coefficients at lag 2 is close to 0 (albeit slightly positive), and the estimates coefficient at lag 4 is also very small (albeit slightly negative). This contrasts with the equivalent lags in the negative model which are negative and positive respectively. The joint hypothesis test for the inclusion of all five returns yields a statistically significant chi-square value of 17.99 ($p < 0.01$) implying that returns jointly impact upon the vice variable. Overall the results yield similar interpretations to those of the negative model - namely, an increase in returns is followed by a decrease in vice terms, and an increase in vice terms is followed by a subsequent decrease in returns. Thus the relationship between vice terms and returns appears to be bidirectional, and there appears to be a feedback mechanism whereby each variable impacts the other.

Table 4.10: Regression of Enron's winsorized returns against negative and vice terms during the period 01/01/1997 through 31/12/2001 ($N = 1,191$ days). The table shows two regression models: one in which returns are regressed against negative counts, and one in which returns are regressed against vice counts. In the negative column the counts of the negative variable at time t serve as the dependent variable. In the vice column the counts of the vice variable at time t serve as the dependent variable. The table shows the estimated coefficients (in basis points) of the lagged winsorized returns model (i.e. β_1 in Equation 4.4). All continuous variables are standardized to z-scores, and the estimated coefficients here are shown in basis points. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals. A dot after the coefficient represents significance at the 10% level. A bold font denotes significance at the 5% level. Bold and italics denotes significance at the 1% level.

	Negative	Vice
Lag_{t-1}	-3.6.	-1.9
Lag_{t-2}	-4.0	0.00
Lag_{t-3}	-4.3	-5.5
Lag_{t-4}	1.2	-0.2
Lag_{t-5}	-0.7	-2.6
$\chi^2(5)[Joint]$	21.64	17.99
MSE	0.382	0.373

Overall, these results tentatively suggest that the relationship between sentiment (ethical terms) and returns is bidirectional. While historical negative and vice terms impact on next-day returns, the past value of returns also impacts on negative and vice terms. This supports the previous literature in Tetlock (2007), which found that the relationship between returns and negative sentiment was bi-directional. This is a promising finding, given that the results in Tetlock (2007) were reported at the index-level, whereas the returns used here are at the firm-level. Additionally this study was constructed using a different textual corpus and market data spanning different time periods (thus encompassing a variety of economic conditions).

4.3.4.4 Exploring the Time-Varying Impact of Sentiment

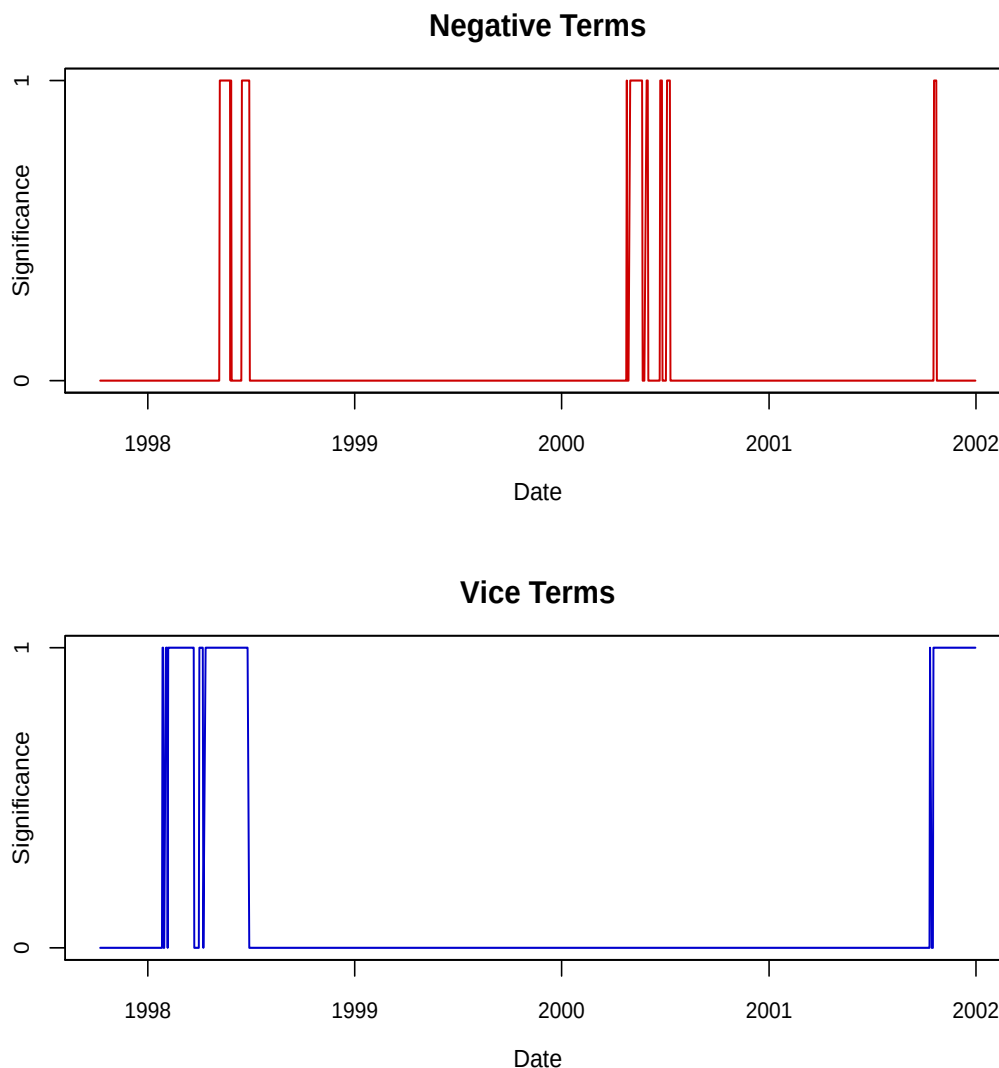
Having identified a relationship between negative/ethical terms and financial returns, I now examine the extent to which this relationship varies over time. To this end I estimate a series of rolling regression models using a 130-day rolling window (comprising approximately 6 months worth of trading days), rolled forward one day at a time. The data and equation fitted in this section are the same as that in Section 4.3.4.2. Specifically, I fit two sets of rolling regressions and explore the impact of negative and vice terms on Enron's winsorized returns.

In Section 4.3.4.2 it was observed that both the negative and vice variables at time $t-1$ had a significant impact on Enron's (winsorized) returns at time t . Accordingly, in each rolling window I compute a t-test on the first lag of the negative and vice variables - that is $negative_{t-1}$ and $vice_{t-1}$. Figure 4.4 shows the significant coefficients for each rolling window. In this figure a value of 1 denotes significance at $p < 0.05$ and a value of 0 denotes $p \geq 0.05$.

Figure 4.4 shows that during certain periods the negative and vice variables are able to account for variability in next-day returns that cannot be accounted for by any of the other financial variables. This supports the previous case study's findings that the impact of negative and vice terms on next day returns is indeed episodic. Specifically, out of the 1,062 rolling models for each equation (Equation 4.4), the first lag of the negative variable ($negative_{t-1}$) is significant in only 52 models (4.90%). Contrarily, the first lag of the vice variable ($vice_{t-1}$) is significant in 147 models (comprising 13.84%). This would suggest that when using a rolling window, the vice terms contain useful information about the returns at time t than is lost when using the broader set of

negative terms. This also suggests that the model parameters in Section 4.3.4.2 vary over time.

Figure 4.4: Rolling regression of negative and vice terms against Enron's returns. The rolling regressions use a 130-day rolling window, rolled forward by one day at a time. A t-test is conducted on the first lag of the negative or vice term in each period. The figures here show the significant estimated regression coefficients: a significance of 1 denotes $p < 0.05$ for the particular rolling window, whereas a significance of 0 denotes $p \geq 0.05$.



Another interesting observation from this figure, is that there is only a small amount of overlap between the two models. That is, of the 147 models in which the vice terms at $t - 1$ were statistically significant, only 25 models overlap with the 52 models in which the negative terms at $t - 1$ were statistically significant. This lack of overlap suggests that the vice terms in the GI dictionary are capturing information about the variation in returns that cannot be captured by using the negative terms. However, the

complete set of negative terms also contains information about returns that cannot be accounted for using only the vice terms.

4.3.5 Summary of the Enron case study

In this case study it was observed that the effect of sentiment was limited in the presence of extreme outliers. After trimming extreme values to cap their influence, both the negative and vice variables have a significant impact on next-day firm-level returns. Unlike in Tetlock (2007), the sentiment variables are not reversed following the initial shock. Rather, the shock from the sentiment variable is persistent and almost all lags exert a negative influence on financial returns. This is perhaps to be expected, given the large decline in Enron's returns over the entire series (on average the returns were -0.34% across the entire time period).

It was also observed that the impact of the negative and vice variables on returns was bi-directional. While the historical negative and vice variables impacted Enron's returns (both jointly and individually at lag $t - 1$), the historical returns were also significant predictors of both the negative and the vice variables. This supports previous literature (e.g. Tetlock (2007), García (2013)), which found that the relationship between negative sentiment and returns was bi-directional, and showed that there was a feedback effect between both variables. However while Tetlock (2007) and García (2013) were constructed at the aggregate (index) level, this case study explored sentiment at the firm-level.

This study also discovered that the impact of negative and vice terms on financial returns was episodic, in that the impact was confined to particular time periods. A set of rolling VAR models (using a 130-day rolling window) revealed that in 147 models (13.84%) the vice terms had a statistically significant impact on next-day returns, in contrast to only 52 models (4.90%) for the negative terms. This finding seems somewhat contradictory to intuition, given that the vice terms correspond to a subset of the negative terms and therefore the negative terms both encompass the vice terms and more. However, the Enron study was chosen deliberately on the basis that it dealt with unethical activity.

One of the lessons learned from the Enron case study was that it is very difficult to conduct VAR analysis using firm-level data. Because news does not occur on every day of the week, it is very difficult to construct a continuous sentiment series with no gaps.

Indeed, this problem is also compounded by the chosen time frame (1997-2001) which pre-dated the digitalization of the historical news articles. Perhaps a better solution would be to conduct an analysis at the aggregate (index) level using a collection of companies from the same industry, and this in fact is the approach adopted in the next case study.

4.4 Case Study 3: Sentiment at the Industry-Level

4.4.1 Introduction

So far this chapter has explored the connection between vice and negative terms in general financial news (the AofM corpus) and in firm-level news about a prominent financial scandal (the Enron case study). Whereas the AofM study used a market-wide index (the DJIA index) for the dependent variable, the Enron study used firm-level returns. There were also differences in the news sources used to derive the sentiment proxy. In the AofM study, the corpus comprised a single daily opinion column, which discussed general market activity of the previous day. In the Enron study, the corpus comprised a wider variety of news sources (both opinion and non-opinion pieces), with news that was written about only a single company. In this case study I decided to investigate the role of sentiment and ethical terms at the industry-level, using both news and financial data from the pharmaceutical industry - an industry where ethics are brought to the forefront of modern business practices.

The pharmaceutical sector deals with ethical issues on a daily basis. Companies from this sector manufacture medication, equipment and provide services that can prevent diseases, ease pain and ultimately save lives. However, companies from this sector are also enormously profitable and it is not uncommon for companies in this sector to have a market capitalization worth hundreds of billions of dollars⁵¹. In an ever-competitive market, and with a desire to maximize profits, these companies occasionally engage in unethical business practices, such as promoting drugs for unauthorized uses (US Department of Justice (2012)), falsifying the results of clinical trials (Wall Street Journal (2015)), extortionate pricing of medication (Fortune Magazine (2015)) and

⁵¹Johnson & Johnson (NYSE: JNJ) and Pfizer (NYSE: PFE) have market caps of 326 and 210 billion respectively as of 04th September 2016. This information was obtained from the Yahoo! Finance website.

many others. Pharmaceutical companies thus engage in both ethical and unethical business practices, which provides a solid motivation for their study. Of course the purpose of this investigation is not to pass judgement on these companies, but to see whether ethical terms, as expressed in news, have any impact on financial markets in the same manner previously ascribed to negative terms.

The work in this section focuses on pharmaceutical companies that engage in the production of pharmaceutical and biopharmaceutical goods. Here the term ‘pharmaceutical’ is used to refer to both of these industries collectively. Most existing research in the literature has focussed on exploring market sentiment at the index level, where indices such as the DJIA or S&P500 are typically used as a proxy for market value. A few studies have focussed on exploring the effect of sentiment at the firm-level, however the sparsity of the data leads researchers to aggregate companies’ returns using methods such as panel regression. In this study, I adopt a mid-way point by analysing firm-level news at the *industry* level. Specifically, I analyse the effect of textual sentiment written about 25 companies listed in the New York Stock Exchange’s ARCA Pharmaceutical Index. Using news about these companies I investigate the impact of both negative and vice terms on the index’s returns, using a series of VAR and rolling VAR models. As in the previous two case studies the system in this thesis is used to analyse the news and compute the frequencies of negative and vice terms.

4.4.2 Data

In this study I explore the impact of news in the pharmaceutical industry, using the NYSE’s ARCA Pharmaceutical Index (Ticker: DRG) as the primary dependent variable of interest. The ARCA pharmaceutical index is ‘*designed to represent a cross section of widely held, highly capitalized companies involved in various phases of the development, production, and marketing of pharmaceuticals.*’ (New York Stock Exchange (2014)). The index is market-capitalization weighted (meaning that the companies’ relative size is taken into consideration in the formation of the index), with the price calculated from the closing price of each listed company at the end of each trading day. In order for a company to be eligible as a component of the ARCA index they must be listed on either the NYSE, NYSE AMEX, NYSE ARCA or NASDAQ stock exchanges. Additionally, they must maintain a market value of at least \$75 million, and have had a trading volume of at least one million shares in each of the last six months.

Thus, each of the companies are of a substantial size, and each company is listed on a major stock market index. Historical prices for the ARCA index were obtained from the Yahoo! finance website (finance.yahoo.com, Accessed: 13 October 2016) for the period 01/01/2007 through 31/12/2015. During the chosen period of investigation, returns of the ARCA pharmaceutical index exhibit an 82.3% correlation with the DJIA index ($r = 60.15$, $p \leq 2.2\text{e-}16$), which provides further motivation for using the ARCA index as a proxy for market value. To the best of my knowledge this is the first study to use this index, rather than a general market index (e.g. DJIA, S&P500) or individual firm-level returns that are typically used in studies on sentiment analysis. The components of the ARCA pharmaceutical index are listed in Table 4.11.

An early decision was made to analyse news at the individual firm-level. While it is possible that industry-level news could affect the pharmaceutical index, it was believed that news about individual components were more likely to influence the index's value. Obtaining this firm-level news was no trivial task, as it proved extremely challenging to find a news source with sufficient daily coverage across the time period of study. Ultimately I decided to use newswire texts, as these were found to be much more frequent than newspaper counterparts. Another advantage of newswire texts is that they are likely to represent the primary sources of information used to produce newspaper articles. As a result the information in newswire texts is likely to be fresher. During this period two newswire sources gave adequate coverage of the chosen pharmaceutical companies. These were PR Newswire and Business Wire, which were also used by Shynkevich et al. (2016), Davis et al. (2012) and Davis & Tama-Sweet (2012).

The corpus was constructed by downloading PR Newswire and Business Wire texts from the LexisNexis database. In this study I searched LexisNexis for all business-related news from PR Newswire and Business Wire for the 25 pharmaceutical companies listed in Table 4.11. To ensure that each text was written about a particular company, as opposed to merely mentioning them in passing, I required that each document contain the company name (or any common abbreviation) within the headline, and at least 3 mentions in the body of the text. To maximize coverage I allowed for short-forms/abbreviations of the names. For instance, for 'Bristol-Myers Squibb' the shortened names of 'Bristol-Myers', 'Bristol Myers' or 'Myers Squibb' were allowed to occur in the headline/body. Similar allowances were made for common abbreviations, such as 'J & J' for 'Johnson & Johnson' and 'GSK' for 'GlaxoSmithKline'. This cri-

terion was based on manual introspection and subsequent refinement of the retrieved texts. The resulting corpus contains 17,701 documents (approximately 13.6 million tokens) spanning the period 01/01/2007 to 31/12/2015. After removal of exact duplicate documents this number reduced to 14,332 documents (approximately 11 million tokens). A breakdown of the corpus composition is given in Table 4.12.

On average there are 18.3 negative terms and 6.7 vice terms per every 1,000 words of text. This suggests that in the corpus the negative terms are approximately three times as frequent as the vice terms, echoing the results found in the previous case studies. The average article length is approximately 770 words. Assuming an average of 252 trading days per year⁵² there are just under 8 articles per trading day over the period 01/01/2007 through 31/12/2015. As a result, the pharmaceutical corpus is much more densely populated than the other two corpora presented in this chapter.

To ensure comparability to existing literature (e.g. Tetlock (2007), García (2013)) and remove the impact of other variables known to impact market returns, volume and volatility estimates are included into the regressions as per the previous case studies. As a measure of volume I use historical NYSE volume, and obtained historic NYSE volume data from the NYSE's 'Transactions, Statistics and Data Library' (<https://www.nyse.com/data/transactions-statistics-data-library>, Accessed: 13 October 2016). For the volatility variable, I used the Chicago Board Option Exchange's (CBOE) Volatility Index, obtaining historical volatility values from Yahoo! Finance (finance.yahoo.com, Accessed: 16 November 2016). The CBOE Volatility Index, more commonly referred to as the 'VIX' index based on its ticker symbol, represents the expected market volatility over the next thirty days, and gives comparable results to the volatility proxy used in Tetlock (2007). Both the volume and volatility proxies are log transformed to remove any heteroscedasticity in the series, and detrended by subtracting a backwards 60 day moving average.

⁵²On average there are 365 days per year, of which 5/7 are working days, corresponding to 261 days. Of these 261 days 9 are US holidays, when the stock market is closed. As a result, there are $261 - 9 = 252$ trading days per year on average.

Table 4.11: Components of the ARCA Pharmaceutical Index as of 25 January 2016. The Yahoo! Finance ticker symbol used to identify a particular stock is shown in the first column, together with the exchange upon which the company trades listed in brackets. The company name is shown in the second column.

Ticker (Exchange)	Company Name
ABBV (NYSE)	AbbVie
ABT (NYSE)	Abbott Laboratories
AGN (NYSE)	Allergan
ALKS (NASDAQ)	Alkermes
AZN (NYSE)	AstraZeneca
BMJ (NYSE)	Bristol-Myers Squibb
ENDP (NASDAQ)	Endo International
GSK (NYSE)	GlaxoSmithKline
JAZZ (NASDAQ)	Jazz Pharmaceuticals
JNJ (NYSE)	Johnson & Johnson
LLY (NYSE)	Eli Lilly
MNK (NYSE)	Mallinckrodt
MRK (NYSE)	Merck
MYL (NASDAQ)	Mylan N.V.
NVO (NYSE)	Novo Nordisk
NVS (NYSE)	Novartis
PFE (NYSE)	Pfizer
PRGO (NYSE)	Perrigo
RDY (NYSE)	Dr. Reddy's Laboratories
SHPG (NASDAQ)	Shire
SNY (NYSE)	Sanofi
TARO (NYSE)	Taro Pharmaceuticals
TEVA (NYSE)	Teva Pharmaceuticals
VRX (NYSE)	Valeant Pharmaceuticals
ZTS (NYSE)	Zoetis

Table 4.12: Composition of the pharmaceutical corpus

Number of	Frequency
Articles	14,332
Days (inc. weekends)	2,500
Negative terms	201,982 (1.83%)
Vice terms	74,007 (0.67%)
Total terms	11,037,521

4.4.3 Methods

During the period of study (01/01/2007 through 31/12/2015) there was a significant period of financial turbulence, caused by the 2007-2008 global financial crisis. This financial crisis led to a period of global economic turn-down, with some economists labelling it as the worst financial crisis since the Great Depression of the 1930s (Temin (2010)). Even the pharmaceutical industry - usually resistant to financial recessions - was affected by the crisis. Figure 4.5 shows a plot of the ARCA Pharmaceutical returns over the period 01/01/2007 to 31/12/2015. From this figure it is clear that there is a major period of volatility during 2008. This is evidenced by the large spikes in returns relative to the rest of the series. Given that the period of study encompasses this major turbulent period, the decision was made to follow the approach in García (2013) by splitting the time series into two distinct periods: a recessionary period (01/12/2007 through 30/06/2009), and an expansionary period (01/07/2009 through 31/12/2015). These periods were determined using the National Bureau of Economic Research's (NBER's) business cycles information (<http://www.nber.org/cycles.html>, Accessed: 13 October 2016).

Table 4.13 gives descriptive statistics of the ARCA index returns across the recessionary, expansionary and overall periods. Here, one can see that the recessionary period is marked by significantly lower returns (an average of -7.7×10^{-4}) relative to the expansionary period (an average of 4.43×10^{-4}). Additionally, the returns during the recessionary exhibit a higher degree of variation than returns during the expansionary period, as evidenced by the standard deviations (1.80×10^{-2} vs. 0.89×10^{-2} respectively). These descriptive statistics suggest that the statistical distributions of

returns are different during the recessionary period compared with the expansionary period, and therefore motivates the decision to split the time series into two distinct periods.

Figure 4.5: Returns of the ARCA pharmaceutical index during the period 01/01/2007 to 31/12/2015. The recessionary period is clearly evident during the 2008-2009 period, as shown by the high spikes in returns during this time frame.

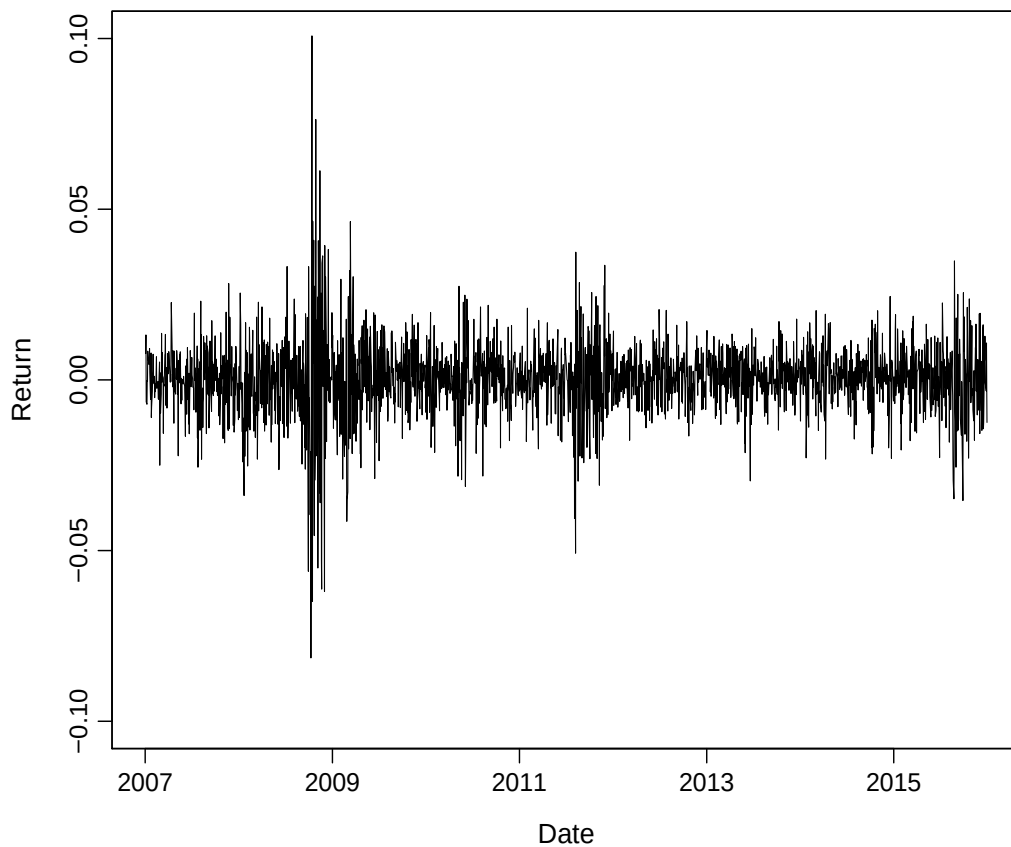


Table 4.13: Descriptive statistics for the ARCA pharmaceutical index. Results shown are calculated using continuously compounded returns for the period 2007-2015. The statistics here were calculated during three separate periods corresponding to an overall period (encompassing both recessionary and expansionary periods), a recessionary period, and an expansionary period. Mean returns are multiplied by 10^4 and standard deviations by 10^2 for ease of readability.

Period	Dates	No. Observations	Mean x 10^4	St. Dev x 10^2
Overall	01/01/2007 to 31/12/2015	2266	1.99	1.10
Recessionary	01/12/2007 to 30/06/2009	397	-7.70	1.80
Expansionary	01/07/2009 to 31/12/2015	1638	4.43	0.89

For both the recessionary and expansionary periods I fit two VAR models of the form:

$$r_t = \beta_0 + \beta_1 L_5 r_t + \beta_2 L_5 Sentiment_t + \beta_3 L_5 Volume_t + \beta_4 L_5 Volatility_t + \beta_5 Calendar_t + \varepsilon_t \quad (4.5)$$

Where:

r_t = returns of the ARCA pharmaceutical index at time t

$Sentiment_t$ = the (standardized) relative frequencies of negative or vice terms

$Volume_t$ = log-transformed and de-trended volume of the NYSE

$Volatility_t$ = log-transformed and de-trended VIX index returns

$Calendar_t$ = January and day-of-the-week dummy variables

β_0 = a constant term

ε_t = an error term, assumed to have a mean of 0

All continuous variables in this equation are standardized into z-scores, by subtracting the mean and dividing by the standard deviation. This ensures that the different variables are standardized to a common unit, that enables for equal comparison of the impact.

One of the assumptions of regression using time series data is that the variables used are stationary. Table 4.14 shows the results of stationarity tests (Augmented Dickey-Fuller) on each of the continuous time series variables used in Equation 4.5. Here one can see that all variables meet the stationarity requirement.

Table 4.14: Augmented Dickey-Fuller (ADF) tests for stationarity of continuous time series variables. Stationarity is tested for variables up to five lags. The ADF test tests for null hypothesis of non-stationary time series, thus p -values < 0.05 are observed to be stationary.

Variable	Dickey-Fuller Statistic	P-Value
Returns (ARCA)	-20.19	< 0.01
Volume (NYSE)	-12.39	< 0.01
Volatility (VIX)	-22.03	< 0.01
Negative	-16.60	< 0.01
Vice	-17.01	< 0.01

Identification of Mistags : The Role of Specialist Dictionaries

The General Inquirer dictionary has occasionally been subject to criticism, and it has been claimed (e.g. Loughran & McDonald (2011)) that some of the dictionary’s negative terms do not have evaluative connotations in certain domains. For instance, the word ‘cancer’ is labelled as both a negative and vice term in the GI dictionary, but would not necessarily be used in an evaluative manner in the pharmaceutical domain (i.e. in discussions of cancer treatment and prevention). In this case study I therefore employ a domain ‘override’ dictionary containing a list of domain-specific (non-evaluative) words and phrases from the pharmaceutical industry. This is essentially a stop-word list which of words and phrases which the system is instructed to ignore. The system first performs a lookup in this override dictionary to see whether a given word or phrase is present: if it is, then the system is instructed to ignore the token, and label it instead as a domain-specific item. The motivation for using the domain override dictionary arose from an examination of the most frequently occurring negative and vice terms in the pharmaceutical corpus⁵³. Here it was observed that a number of the most frequently occurring evaluative terms had literal (i.e. domain-specific) meanings, and so should not be labelled as evaluative or ethical by the system. This section examines the most frequently occurring negative/vice terms in the pharmaceutical corpus, before discussing the composition of the domain override dictionary used in this case study.

Table 4.15 shows the top 20 most frequent negative word senses across the entire pharmaceutical corpus. Collectively, these top 20 negative terms account for 48.41% of all negative terms in the corpus, suggesting that any potential mistagging in this list could have a drastic impact on the results. An implication of this is that remedying any mistags in the top 20 list, will have fixed the mistags for almost half of all negative words. I define a mistag as a word that is not typically evaluative in the target domain. For the pharmaceutical domain this includes any healthcare-related words that may not be typically evaluative. Some obvious candidates from the list are the words

⁵³This was a manual process, although one would conjecture that this process could potentially be automated - for instance by comparing whether there are certain evaluative terms that occur in a corpus with a higher frequency than would normally be expected. This is an area left for future research, and could potentially be incorporated into future versions of the system.

‘disease’, ‘cancer’, ‘disorder’ and ‘infection’, which would not necessarily be negative in a pharmaceutical context. The terms ‘chronic’ and ‘severe’ could also possibly be considered mistags, depending on the context. Often these terms appear as part of multi-word disease names or symptoms (such as ‘chronic hepatitis’ which occurs 580 times in the corpus, or ‘chronic disease’ which occurs 282 times). Accordingly, any domain dictionary should incorporate these multi-word expressions.

Table 4.16 shows the top 20 vice word senses in the corpus for the same period. As with the negative terms, the top 20 most frequent vice terms account for the bulk of all vice terms. Indeed, the effect is even more severe for the vice terms, as the top 20 terms alone account for 71.23% of all vice terms in the corpus. The top 20 vice terms also contain a number of terms that would not necessarily be considered evaluative (ethical) in the pharmaceutical domain. The obvious candidates are ‘disease’, ‘cancer’, ‘symptom’, ‘headache’ and ‘fracture’, but again the terms ‘disorder’, ‘chronic’ and ‘severe’ could potentially be non-evaluative if occurring as part of larger multi-word expressions. One might wonder why these words are labelled as vice terms in the dictionary in the first place. It is most likely that this is attributable to a limitation in the coverage of the GI’s disambiguation rules⁵⁴. Kelly & Stone (1975) created disambiguation procedures only for the most frequent word types in their corpus of General English, however there may well be some remaining low-frequency words in the dictionary that are ambiguous. These are presumably tagged with multiple categories to cover all possible senses, and because these words usually occur with such low frequency, the authors probably decided that this was an acceptable margin of error. So while in certain contexts these words might be considered vice terms, they are clearly not being used in a vice sense here. As their senses are aggregated together in the GI dictionary, and because the terms occur here with much higher frequency due to the domain, this has naturally inflated the number of vice terms.

⁵⁴Another possible suggestion is that some of these words might have a psychological or sociological meaning that is not typically used in general English.

Table 4.15: Top 20 most frequent negative words senses in the pharmaceutical corpus (spanning the period 01/01/2007 to 31/12/2015). Note that the number following the # corresponds to the corresponding sense number in the GI dictionary.

Word Sense	Frequency	Frequency Per 100,000 words
Disease	12,392	112
Cost#1	11,992	109
Expense#1	11,613	105
Cancer	8,576	78
Loss	6,985	63
Exclude	4,781	43
Adverse	4,575	41
Liability	3,611	33
Against	3,347	30
Need#1	3,315	30
Lower#1	3,251	29
Need#2	3,204	29
Pain	2,810	25
Decrease#1	2,793	25
Disorder	2,709	25
Chronic	2,670	24
Infection	2,451	22
Drive#2	2,321	21
Close#2	2,191	20
Severe	2,191	20
Total	117,699	1066

Thus there are a number of evaluative and ethical words in the GI dictionary that are not typically evaluative or ethical in the pharmaceutical corpus. To prevent these mistags from unduly influencing the regression equations, I decided to use a domain-specific override dictionary whose entries take precedence over those in the GI dictionary. The system first checks whether a term (single or multi-word) is in the override dictionary. If it is in the dictionary then the system ignores it. If not, then the system proceeds to look-up the term in the original GI dictionary and add the associated tags to the document. This allows the user to tailor the system for any domain, by simply

using an appropriate override dictionary.

Table 4.16: Top 20 most frequent vice senses in the pharmaceutical corpus (spanning the period 01/01/2007 to 31/12/2015). Note that the number following the # corresponds to the corresponding entry in the GI dictionary.

Word Sense	Frequency	Frequency Per 100,000 words
Disease	12,392	112
Cancer	8,576	78
Loss	6,985	63
Adverse	4,575	41
Disorder	2,709	25
Chronic	2,670	24
Severe	2,191	20
Symptom	2,087	19
Negative	2,036	18
Error	1,236	11
Failure	1,084	10
Unfavorable	1,009	9
Mistake#1	771	7
Headache	725	7
Rat	690	6
Fracture	657	6
Mean#2	599	5
Damage#1	587	5
Difficult	583	5
Limitation	551	5
Total	52,713	478

While dictionary construction is a valid research topic in itself it was not the core focus of this thesis. There are a multitude of techniques devoted to dictionary construction, however I do not discuss these here⁵⁵. Instead, I used two domain-specific

⁵⁵I debated, for instance, using WordNet to derive a list of all the hyponyms of the ‘ill health’ and ‘symptom’ categories, however the coverage of this was inadequate and the resource was not constructed by domain experts.

dictionaries that were compiled by experts, namely a Human Disease Ontology and a Symptom Ontology both from the Institute for Genome Sciences at the University of Maryland (Schriml et al. (2012), Kibbe et al. (2014) and Osborne et al. (2009)). The Disease Ontology comprised 9,180 tokens and the Symptom Ontology comprised 939 tokens. The two ontologies were combined into a single override dictionary of 10,119 tokens in a format that could be processed by the system. Over 88% of terms in this dictionary were multi-word expressions, consisting of two or more words (words here are defined as a sequence of alphanumeric characters separated by white space). This use of multi-word expressions partially captures contextual information, akin to the multi-word idiom routines used by the GI disambiguation rules. For instance, consider the token ‘chronic’ which appears amongst the top 20 negative and vice terms in Tables 4.15 and 4.16. While the override dictionary does not contain the token ‘chronic’ in isolation, there are 88 multi-word expressions containing this word, such as ‘chronic fatigue syndrome’ and ‘chronic cholangitis’. Thus if ‘chronic’ appears in isolation it will be tagged as negative or vice, but if it appears as one of the multi-word expressions in the override dictionary it will be ignored. The top 30 most frequent words in the override dictionary are shown in Table 4.17, with the complete list available for download on the authors website (<http://disease-ontology.org>, Accessed: 14 December 2016). While it is acknowledged that some of these terms (e.g. ‘cancer’) could possibly occur in a metaphorical sense, and that this is a potential limitation of this dictionary, the most frequent senses generally appear to be used in a non-metaphorical manner. For instance, the word ‘cancer’ occurs 9,825 times in the pharmaceutical corpus, equating to 11.23 occurrences per every 10,000 words. By comparison, ‘cancer’ is labelled as a metaphor in the news section of the VU Amsterdam Metaphor Corpus (Vu et al. (2012)) only 3 times, equating to 0.66 times per 10,000 words⁵⁶. Thus even if some of the occurrences of ‘cancer’ (and similar terms) are used in a metaphorical sense, it is unlikely to have a significant impact upon the results. In the remainder of this study all frequencies of negative and vice terms are computed using the domain override dictionary (in conjunction with the original GI dictionary) unless stated otherwise.

⁵⁶Additionally, the news section in the VU Amsterdam Metaphor Corpus is of the general language variety, as opposed to the business-oriented pharmaceutical news in this case study. One would expect the token ‘cancer’ to be used in a more restrictive sense in the news in this study, and more broadly in the Metaphor Corpus.

Table 4.17: Top 30 most frequent terms in the pharmaceutical override dictionary. These terms, ranked from most-least frequent, equate to 62.34% of all override terms in the pharmaceutical corpus.

Term	Frequency	Frequency Per 100,000 words
Disease	10,491	95
Cancer	9,451	86
Pain	3,151	29
Symptom	2,468	22
Hepatitis	2,111	19
Influenza	1,147	10
Asthma	1,009	9
Syndrome	825	7
Osteoporosis	769	7
Headache	755	7
Psoriasis	714	6
Rheumatoid Arthritis	680	6
Nausea	648	6
Depression	636	6
Schizophrenia	613	6
Melanoma	603	5
Diarrhea	600	5
Obesity	597	5
Stroke	584	5
Multiple Sclerosis	576	5
Migraine	567	5
Inflammation	518	5
Rash	481	4
Vascular Disease	476	4
Bleeding	468	4
Fatigue	463	4
Myopathy	456	4
Noma	426	4
Leukemia	423	4
Vomiting	417	4
Total	43,123	391

4.4.4 Experimental Results

4.4.4.1 Impact of Negative and Vice Terms on ARCA Returns

In this section I explore the impact of the GI's negative and vice terms on returns of the ARCA pharmaceutical index over the period 2007 to 2015. Because this time period encompassed a major recessionary period, I split the data into two smaller time periods: a recessionary period from 01/12/2007 through 30/06/2009, and an expansionary period from 01/07/2009 through 31/12/2015. In each period I fit two VAR models of the form in Equation 4.5: one in which the *Sentiment* variable denotes the frequency of negative terms and one which it denotes the frequency of vice terms. Thus in total, this section reports the results of 4 regression equations. Because the main focus of this thesis is on the impact of evaluative and ethical terms, I report only the results of these variables. Specifically, I focus on the impact of β_2 (Equation 4.5), which describes the impact of lagged negative (vice) terms on the dependent variable. The results of the four regression equations, are shown in Table 4.18, with Panel A denoting the results of the regressions during the recessionary period and Panel B denoting the results of the regressions during the expansionary period.

In the recessionary period (Panel A) the estimated regression coefficients are generally of a larger absolute magnitude than those during the expansionary period (Panel B). In addition, the recessionary period does not follow the same pattern of a negative coefficient at lag 1 that is observed both in the expansionary period and also in the previous case studies. For the negative model (Panel A) the coefficient at lag 4 is both positive and statistically significant ($p < 0.05$), suggesting a delayed impact of negative terms on index returns. The joint hypothesis test for inclusion of all 5 lags of negative sentiment yields a statistically significant χ^2 value of 11.81 ($p < 0.05$), suggesting that the negative terms have a joint impact on the ARCA index returns. In terms of the vice model the coefficient at lag 2 is both negative and highly significant ($p < 0.01$) implying that a unit increase in the frequency of vice terms two days prior has an average -16.7 basis point decrease in returns on day t . This effect is subsequently reversed over subsequent days. The coefficient at lag 4 is also positive and marginally significant ($p < 0.10$) further supporting the notion of a reversal.

Table 4.18: Regression of negative and vice terms against ARCA pharmaceutical returns. The results here show the estimated coefficients of lags 1-5 of the *Sentiment* variable from Equation 4.5. The negative columns denote models computed using the frequency of negative terms, and the vice columns denotes models computed using the frequency of vice terms. Panel A represents the two regressions during the recessionary period (01/12/2007 to 30/06/2009) comprising 331 observations. Panel B represents the VAR regressions in the expansion period (01/07/2009 to 31/12/2015) comprising 1,572 observations. The estimated coefficients are shown in basis points. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals up to five lags. A bold coefficient denotes significance at the 5% level. A dot after the coefficient denotes significance at the 10% level.

Panel A: Recessionary Period		
	Negative	Vice
Lag_{t-1}	5.6	1.8
Lag_{t-2}	-9.6	-16.7
Lag_{t-3}	4.1	6.9
Lag_{t-4}	14.3	11.0.
Lag_{t-5}	-0.1	-0.6
$\chi^2(5)[Joint]$	11.81	8.24
MSE	0.954	0.952
Panel B: Expansionary Period		
	Negative	Vice
Lag_{t-1}	-4.5.	-5.7
Lag_{t-2}	1.1	2.3
Lag_{t-3}	4.3.	1.8
Lag_{t-4}	-3.4	-2.2
Lag_{t-5}	1.5	3.5
$\chi^2(5)[Joint]$	6.22	9.32
MSE	0.990	0.989

In Panel B, the estimates of the negative and vice VAR equations are shown for the models fitted during the expansionary period. Here both variables have an initial negative impact on returns, which is subsequently reversed throughout the rest of the week. For the negative model lags 1 and 3 are marginally significant at $p < 0.10$, with the coefficients generally exhibiting similar signs and magnitudes to those in the first case study. The chi-square test for the joint impact of lags 1-5 of the negative variable is not statistically significant, even at the 10% level. In terms of the vice model the

vice terms at lag 1 yields a statistically significant coefficient of -5.7. This implies that during the expansionary period a one unit increase in vice terms leads to a -5.7 basis point decrease in ARCA returns. This result is consistent with the results in the Enron case study. The joint hypothesis test for inclusion of all 5 lags of the vice variable is not statistically significant at conventional levels, though exhibits marginal significance at $p < 0.10$. Overall these results suggest there is no simultaneous impact of all lags on the returns.

4.4.4.2 Consistency Checks

In the previous section I introduced the notion of a domain override dictionary, containing a list of domain-specific words and phrases that were not typically evaluative in the pharmaceutical domain. In this section I present the equivalent results computed without the domain override dictionary. Consequently, I fit four additional regression models using the same data, equations and time periods previously described. The results of these regression models are shown in Table 4.19.

In general, the estimated coefficients in the recessionary period are higher than the equivalent values in the expansionary period. However, none of the estimated coefficients for either the vice or the negative models are statistically significant at conventional levels ($p < 0.05$). Joint hypothesis tests for the joint impact of all 5 lags of negative and vice variables are also not statistically significant in either the recessionary or expansionary periods.

However, while the negative and vice terms do not significantly impact the ARCA returns, some interesting observations can still be drawn. The first is that the signs of the coefficients for the negative model are all the same as the signs of the coefficients for the vice model, suggesting that the vice and negative terms have similar impacts upon the pharmaceutical index. The second interesting point to note is that the first lag of the negative model in Panel B exhibits the same -4.4 basis point magnitude observed in the first case study. While this coefficient is not significant at the conventional levels (although marginally significant at $p < 0.10$), it is encouraging to note the similar pattern to Table 4.2 even using a different textual data, market data and time period. Note that this negative coefficient on the first lag was also observed in the firm-level (Enron) case study.

Overall it appears that the results without the domain override dictionary underes-

timate the ‘true’ effect of the negative sentiment and ethical terms upon returns. This implies that some of the domain-specific pharmaceutical vocabulary was inadvertently affecting the sentiment and ethical proxies. It also provides support for using the override dictionary, and demonstrates that the regression models can be sensitive to the choice of input lexicon.

Table 4.19: Regression of negative and vice terms against ARCA pharmaceutical returns without the domain override dictionary. The results here show the estimated coefficients of lags 1-5 of the *Sentiment* variable from Equation 4.5. The negative columns denote models computed using the frequency of negative terms, and the vice columns denotes models computed using the frequency of vice terms. Panel A represents the two regressions during the recessionary period (01/12/2007 to 30/06/2009) comprising 331 observations. Panel B represents the VAR regressions in the expansionary period (01/07/2009 to 31/12/2015) comprising 1,572 observations. The estimated coefficients are shown in basis points. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals up to five lags. A bold coefficient denotes significance at the 5% level. A dot after the coefficient denotes significance at the 10% level.

Panel A: Recessionary Period		
	Negative	Vice
Lag_{t-1}	7.0	4.9
Lag_{t-2}	-7.0	-11.5.
Lag_{t-3}	7.5	9.5
Lag_{t-4}	9.4	0.3
Lag_{t-5}	-0.8	-1.0
$\chi^2(5)[Joint]$	6.69	5.38
MSE	0.963	0.967
Panel B: Expansionary Period		
	Negative	Vice
Lag_{t-1}	-4.4.	-4.0
Lag_{t-2}	1.6	0.8
Lag_{t-3}	3.8	1.7
Lag_{t-4}	-2.4	-0.6
Lag_{t-5}	1.3	2.4
$\chi^2(5)[Joint]$	6.14	3.98
MSE	0.991	0.992

4.4.4.3 The Impact of Returns on Sentiment

Having established that both negative and vice terms impact returns, it is now prudent to look at the reverse effect. The motivation for this is to establish whether the observed relationship is bidirectional. It may be the case that there is a feedback effect whereby sentiment impacts returns, which in turn impact sentiment. This might happen, for instance, where news is reporting both on information that has already been compounded into the asset price, and on information which has yet to be compounded. In this section I therefore investigate the effect of returns on negative and vice terms. Specifically, I investigate the impact of the ARCA pharmaceutical returns in estimating the frequency of negative and vice terms from the pharmaceutical corpus. To this end, the following model is proposed:

$$\begin{aligned} Sentiment_t = & \beta_0 + \beta_1 L_5 ARCA_t + \beta_2 L_5 Volume_t + \beta_3 L_5 Volatility_t + \\ & \beta_4 Calendar_t + \beta_6 L_5 Sentiment_t + \varepsilon_t \end{aligned} \quad (4.6)$$

Where $Sentiment_t$ denotes the frequency of negative or vice terms at time t , and the other variables are as described in the previous section. This equation is a reversal of the relationship established in Equation 4.5. Table 4.20 shows the estimated coefficients of β_1 , which represent the impact of past ARCA Pharmaceutical returns on the sentiment proxy for both the negative and vice measures. Note that the domain override dictionary is used in this section as per section 4.4.4.1.

The results suggest that ARCA returns are not statistically significant predictors of negative or vice terms neither in the recessionary nor expansionary period. None of the individual coefficients in the model are statistically significant at the conventional levels and the joint hypotheses tests also yield statistically insignificant results. The coefficients calculated during the recessionary period are slightly stronger than during the expansionary period, however are not statistically significant in either period.

In addition to returns failing to impact negative or vice terms, it was also observed that the coefficients for the other variables (β_2 for *Volume*, β_3 for *Volatility*) were also statistically insignificant either at the individual level or the joint level. This suggests that neither volume nor volatility impact sentiment (negative or vice terms) as well as a failure of returns to impact sentiment. Thus while the vice terms impacted returns, the reverse relationship was not found.

Table 4.20: Regression of ARCA pharmaceutical returns against negative and vice terms. Panel A represents the recessionary period (01/12/2007 to 30/06/2009) comprising 331 observations. Panel B represents the expansionary period (01/07/2009 to 31/12/2015) comprising 1,572 observations. The table shows estimations of the of the coefficient β_1 in Equation 4.6. The negative column denotes the impact of lagged ARCA returns on negative terms, and the vice column denotes the impact of lagged ARCA returns on vice terms. Estimated coefficients are shown in basis points. Newey-West robust standard errors are used to account for heteroscedasticity and autocorrelation in the residuals up to five lags. Coefficients in bold and italics denote significance at the 1% level. Coefficient in bold only denote significance at the 5% level. Coefficients with a dot suffix (.) denote significance at the 10% level.

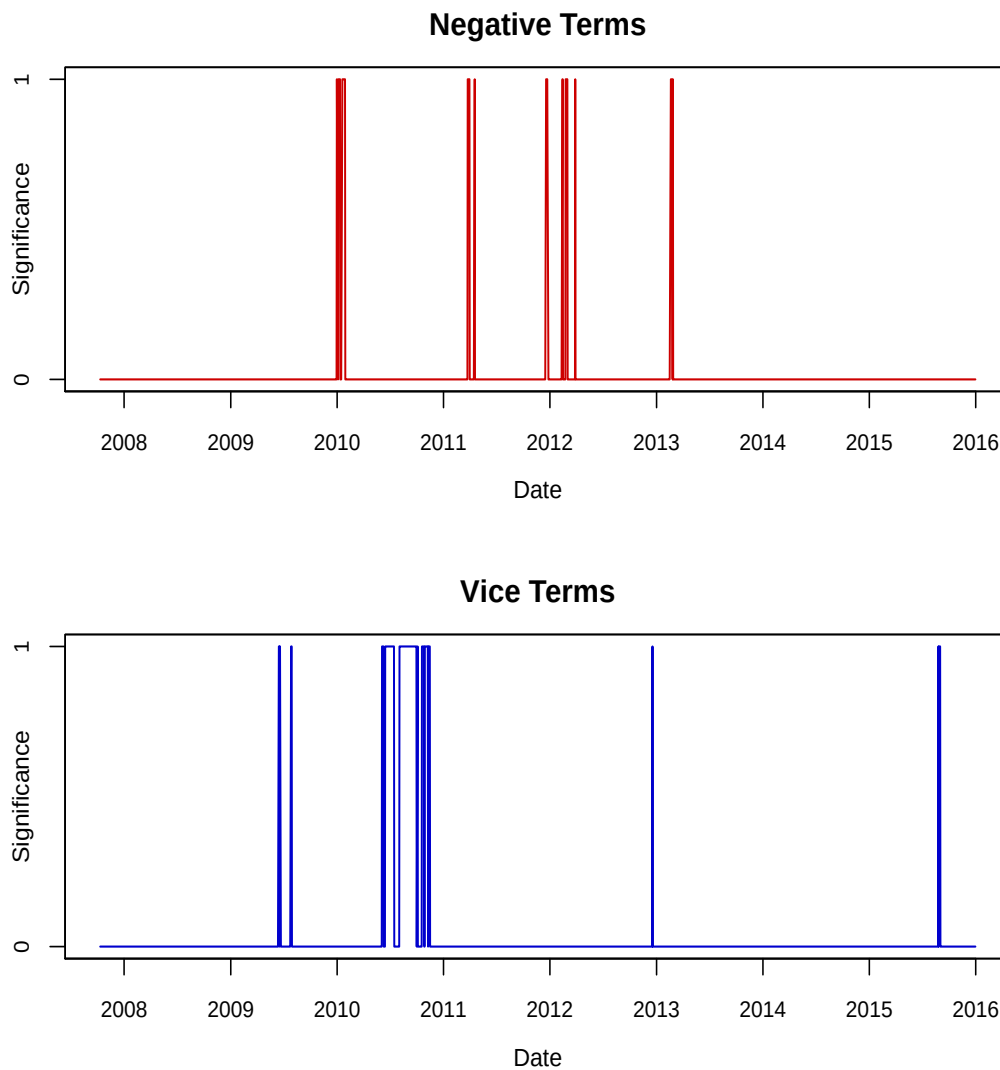
Panel A: Recessionary Period		
	Negative	Vice
$ARCA_{t-1}$	5.0	-0.3
$ARCA_{t-2}$	-4.3	-4.8
$ARCA_{t-3}$	-3.9	-11.1
$ARCA_{t-4}$	10.4	6.2
$ARCA_{t-5}$	0.0	-2.4
$\chi^2(5)[Joint]$	3.13	5.32
MSE	0.857	0.781
Panel B: Expansionary Period		
	Negative	Vice
$ARCA_{t-1}$	0.6	2.0
$ARCA_{t-2}$	-1.1	-1.2
$ARCA_{t-3}$	-0.7	2.6
$ARCA_{t-4}$	3.3	-3.7
$ARCA_{t-5}$	0.1	-0.2
$\chi^2(5)[Joint]$	1.22	2.02
MSE	0.970	0.970

4.4.4.4 Exploring the Time-Varying Impact of Sentiment on Returns

In the previous section it was established that negative and vice terms do impact upon ARCA Pharmaceutical returns, but that the effect is different in the recessionary period to the expansionary period. To further explore this time-varying impact of the derived sentiment proxies, in this section I perform a series of rolling window regressions over the entire sample period (01/01/2007 to 31/12/2015). The results calculated here use

the override dictionary previously discussed in Section 4.4.4.1. The model is thus the same as the previous model (Equation 4.5), except that two rolling VAR models are fitted with a 130-day rolling window (corresponding to approximately 6 months of trading days). Note that the findings with other rolling window sizes (e.g. 250 days) yield similar conclusions.

Figure 4.6: Rolling regression for negative and vice terms. The data used comprises Pharmaceutical newswire texts as previously described. The rolling regression uses a 130-day rolling window, moving the window forward by one day at a time. A t-test is conducted on the first lag of the negative or vice term in each period. The figures here show the significant estimated regression coefficients: a significance of 1 denotes $p < 0.05$ for the particular rolling window, whereas a significance of 0 denotes $p \geq 0.05$.



Using a rolling window width of 130 days resulted in 2,071 models computed using the negative terms, and 2,071 rolling models computed using the vice terms. In each window I calculate the significance of the coefficient at lag $t - 1$, to check whether the prior day's news had any impact on next day returns. The results are presented in Figure 4.6.

Overall, there are 93 models (4.49%) in which the first lag of the negative variable has a statistically significant impact ($p < 0.05$) on next day returns. Conversely, for the vice models there are 98 models (4.73%) in which the first lag of the vice variable has a statistically significant impact ($p < 0.05$) on next day returns. This suggests that both the impact of negative terms and vice terms is indeed episodic, and that the model parameters fitted in the previous sections in this case study vary over time. However, as in the previous two case studies there are a relatively small number of periods in which the negative and vice models overlap ($N=33$). This suggests that while the vice terms appear to be substitutable for negative terms over the entire period, there are certain periods in which the vice terms contain useful information about returns that is lost using the broader subset of negative terms. Conversely, there are also periods in which negative terms contain useful information about the returns that is not available when using only the vice variable in isolation.

4.4.5 Summary of Industry-Level Case Study

In this case study I explored the impact of negative and vice terms at the industry-level. The industry chosen was the pharmaceutical industry where issues of ethical conduct are brought to the forefront of modern business practices. To this end I constructed a corpus of 14,000 documents (11 million tokens) about each of the 25 companies in the NYSE's ARCA pharmaceutical index. I then conducted a series of regressions to explore the impact of negative and vice terms in this corpus on ARCA pharmaceutical returns.

The first key finding is that periods of financial turmoil (viz. a viz. the recessionary period) led to significantly different model estimations than periods of relative tranquillity. In general the absolute magnitudes of the coefficients during the recessionary period were larger than during the non-recessionary period, however the signs and significance levels were very different from those calculated during the non-recessionary period. It may be the case that the volatility in returns during the period was not ade-

quately captured by the volatility proxy. Thus a possible area for future research would be to look at alternative volatility proxies during recessionary periods, such as the use of ARCH or GARCH models which are better designed for dealing with volatility that varies over time.

The second key finding was that words and phrases in the original (unmodified) GI dictionary were not accurately measuring the effect of sentiment and ethical terms in the pharmaceutical domain. A number of the most frequent negative and vice terms in the GI dictionary (such as ‘cancer’, ‘symptom’ and ‘disease’), did not have evaluative connotations in the pharmaceutical domain. Consequently, as a remedial solution I introduced the notion of a domain-override dictionary containing a list of common words and phrases in the pharmaceutical domain that the system should ignore.

The third key finding was that the results in the expansionary period broadly resembled the results in the previous case studies. Although neither the negative nor vice terms had a joint impact on returns, the estimated coefficient at lag 1 was statistically significant for the vice variable ($p < 0.05$) and marginally significant for the negative variable ($p < 0.10$). Interestingly, the estimated coefficient for the negative variable at lag 1 was -4.5, which was very similar to the value of -4.4 observed in Tetlock (2007). Overall the coefficients for the negative and vice variables are broadly similar in terms of the signs and magnitudes. This supports the notion that vice terms can be used to derive an investor sentiment proxy in the pharmaceutical sector.

The final key finding from this study is that the negative and vice proxies had a time-varying impact on financial returns. A series of rolling regression models were computed using a 130-day rolling window, which was rolled forward one day at a time over consecutive time periods. A new regression model was computed in each rolling window, and the significance of the coefficient at lag 1 was noted. Overall the first lag of the negative variable was found to be a statistically significant ($p < 0.05$) predictor of financial returns in 93 models (4.49%), and the first lag of the vice variable was found to be statistically significant in 98 models (4.75%). Interestingly, there were only 33 periods in which the significant models overlap, suggesting that both the negative and the vice models contain some information about financial returns that cannot be captured by only one of the variables. The notion of this time-varying effect concurs with the results found in the previous two case studies.

4.5 Summary

In the previous chapter I described the sentiment analysis system that was developed during this research. This system takes unstructured textual data as input and converts it into a quantitative time series format that can be used in an econometric model. In the present chapter I tested the system's functionality using three case studies, and derived sentiment and ethical proxies from the corpus constructed in each study. I then examined the effect of the derived sentiment and ethical proxies at different levels of market granularities - namely the market-level, industry-level and firm-level. In each study I conducted a series of VAR and rolling VAR models to explore the impact of negative (evaluative) and vice (ethical) terms on financial returns. Table 4.21 summarises the key VAR results in which returns are regressed against negative and ethical terms, and Table 4.22 summarises the key rolling VAR models.

Table 4.21: Summary of the key VAR models for all three case studies. Row 1 corresponds to the Market-Level study, regressing negative and vice terms against DJIA returns (Equation 4.1) during the period 01/01/1984 - 17/09/1999. Row 2 corresponds to the Firm-Level study, regressing negative and vice terms against Enron's returns (Equation 4.3) during the period 01/01/1997 to 31/12/2001. Rows 3a and 3b correspond to the Industry-Level study, regressing negative and vice terms against the DJIA returns (Equation 4.5) - here row 3a corresponds to the recessionary period (01/12/2007 - 30/06/2009) and row 3b corresponds to the non-recessionary period (01/07/2009 - 31/12/2015). The column marked 'N' denotes the number of observations in each model. The columns Negative and Vice denote a row vector of regression coefficients corresponding directly to the coefficients listed in Tables 4.4, 4.8 and 4.18. Coefficients with a dot (.) suffix are significant at the 10% level. Coefficients marked in bold are significant at the 5% level. Coefficients in bold and italics are significant at the 1% level.

Case Study		N	Negative (Lags 1-5)	Vice (Lags 1-5)
(1)	Market-Level	3,709	[- 4.4 , 2.3, -2.4, 3.8 , 2.9]	[-1.6, -0.3, -2.3, 1.1, 0.5]
(2)	Firm-Level	1,191	[- 10.1 , -7.2, -2.4, - 10.6 , -5.8]	[- 13.4 , -4.7, -4.1, -8.8., -2.6]
(3a)	Industry-Level (Recession)	331	[5.6, -9.6, 4.1, 14.3 , -0.1]	[1.8, - 16.7 , 6.9, 11.0., -0.6]
(3b)	Industry-Level (Non-Recession)	1,572	[-4.5., 1.1, 4.3., -3.4, 1.5]	[- 5.7 , 2.3, 1.8, -2.2, 3.5]

Table 4.22: Summary of rolling VAR models for all three case studies. In these models negative/vice terms are regressed against financial returns using a rolling window of 130 days which was rolled forward one day at a time. Row 1 corresponds to the rolling VAR results from the market-level study (Section 4.2.5.3). Row 2 corresponds to the rolling VAR results from the firm-level study (Section 4.3.4.4). Row 3 corresponds to the rolling VAR results from the industry-level study (Section 4.4.4.4). The column marked ‘N’ denotes the total number of rolling VAR models, and the columns marked ‘Sig. Neg’ and ‘Sig. Vice’ denote the total rolling window periods in which lag 1 of the negative/vice variable was statistically significant ($p < 0.05$). The final column, marked ‘Overlap’, denotes the number of models in which both the first lag of negative and the first lag of vice terms were statistically significant ($p < 0.05$).

Case Study	N	Sig. Neg	Sig. Vice	Overlap
(1) Market-Level	3,558	259 (7.29%)	211 (5.93%)	0
(2) Firm-Level	1,062	52 (4.90%)	147 (13.84%)	25
(3) Industry-Level	2,071	93 (4.49%)	98 (4.73%)	33

In the first case study (Row 1 of Tables 4.21 and 4.22) I explored the effect of evaluative and ethical terms on financial returns at the market-level, using returns of the DJIA as a proxy for market-value. Evaluative and ethical proxies were derived from a daily Wall Street Journal opinion column over a 9 year period (1999-2007). While the negative terms had a statistically significant impact on financial returns over the entire period, the vice terms did not (either at the joint level or the individual coefficient level). However, while the vice terms did not appear to have any impact on returns over the course of the entire time period, they did appear to have some impact on financial returns during certain time periods. A series of rolling VAR models (Table 4.22) revealed that both the negative and vice terms had a significant impact on next-day returns during certain time windows, and that both variables contained information about returns was lost when using only one of the variables in isolation. This suggests that while the vice terms were not directly substitutable for the negative terms, during certain periods they did contain some information about returns that was lost using the complete set of negative terms.

In the second case study (Row 2 of Tables 4.21 and 4.22) I explored the effect of general news on firm-level returns using the former Enron corporation as a case study. Unlike the first case study, which used only one opinion-column from a single source, the Enron case study used more general news from a wider variety of sources, spanning the period 01/01/1997 through 31/12/2001. In addition, the news was focussed on

only one particular company, rather than the market as a whole. After winsorizing returns to account for outliers, it was discovered that both negative and vice terms had an effect on Enron's returns, at the individual coefficient level and also at the joint level. This suggests that an increase in negative/vice terms led to a subsequent decrease in Enron's returns. The relationship between returns and sentiment was also found to be bi-directional, in that an increase (decrease) in returns was followed by a decrease (increase) in negative terms, and an increase (decrease) in negative terms was followed by a decrease (increase) in financial returns. Thus a feedback phenomenon appears to exist whereby returns influence sentiment which in turn impacts on returns. Finally, the rolling VAR models (Row 2 of Table 4.22) suggest that the strength of the relationship between negative/vice terms and financial returns varied over time, and that the effect of news on financial returns was confined to particular time periods.

In the third case study (Rows 3a and 3b of Tables 4.21 and 4.22) I examined the effect of sentiment and ethical terms on financial returns at the industry-level, focussing specifically on the pharmaceutical industry. Two major newswire sources were used as an information source, and the impact of evaluative and ethical terms was explored against the NYSE's ARCA pharmaceutical index during the period 01/01/2007 through 31/12/2015. This encompassed a major period of economic turmoil (the recessionary period in 2007-2009) as well as a period of relative tranquillity. The regression results and descriptive statistics during these two periods suggest that the two time periods should be treated independently. During the course of this case study it transpired that a number of the ethical and evaluative terms were not typically used in an ethical or evaluative sense in the pharmaceutical domain. Accordingly, I incorporated a domain override dictionary, containing a list of domain-specific lexical items that the system should ignore. As in the two previous studies the relationship between negative/vice terms and financial returns was found to vary over time. Additionally both the negative and vice models ultimately contained useful information about financial returns that could not be captured using only one of the variables in isolation.

Overall the signs of the model coefficients in each of the three case studies are generally the same for the negative terms as they are for the vice terms. This would seem to suggest that the two variables are generally measuring the same construct. However, the significance levels of the negative and vice terms varies somewhat between the studies: while the negative terms appear to have more of an impact on returns in

market-level study, the vice terms appear to have a larger impact in the industry-level study and both terms have an impact in the firm-level study. With the exception of the recessionary period in the industry-level study it is generally lag 1 of the negative/vice variable that has a significant impact on next-day returns. This suggests that the effect of news is promptly incorporated into the share price, although there does appear to be a somewhat persistent effect in the firm-level study. The coefficient at lag 1 is also generally negative, suggesting that an increase in negative/vice terms leads to a decrease in subsequent returns and, conversely, that a decrease in negative/vice terms leads to an increase in returns. This is consistent with intuition: negative news leads to a decrease in share price. The strength of the relationship between negative/ethical terms and financial returns was also found to vary over time. A series of rolling regressions were computed for each case study in which it was revealed that negative/vice terms had a stronger impact on financial returns during certain time periods, suggesting that there were certain periods in which the markets were inefficient. However, while the negative and vice terms generally seem to be measuring a similar construct, there are certain periods in which the negative or vice terms are able to account for variation in returns that cannot be accounted for using only the other variable in isolation.

Chapter 5

Conclusion

The literature on behavioural finance argues that humans are subject to emotional and psychological biases, that can cause them to act in irrational ways. These emotional responses can cause investors to overvalue or undervalue a given asset beyond its ‘true’ (fundamental) value. By analysing news and other information sources, it may be possible to glean an insight into investors’ thoughts and feelings at an aggregate level. This can be used to determine whether investors perceive an asset’s current market value to be overpriced or underpriced.

Previous approaches to this problem tend to analyse the affective language in text. It is believed that affect functions as a proxy for investor sentiment, describing the overall bullishness or bearishness of a market at a particular point in time. The majority of existing approaches are dictionary-based, using a pre-compiled dictionary of sentiment terms. Typically the affect categories of negative or positive are used, and therefore sentiment analysis amounts to counting the number of negative or positive terms in a given collection of texts. However, a large subset of affective terms in dictionaries often have ethical undertones. For instance, affective terms such as ‘harm’, ‘deception’ and ‘selfish’ invoke connotations of unethical behaviour. The main objective of this thesis was to determine if this ethical subset could be used to derive a proxy of investor sentiment.

The motivation for investigating the relationship between ethical terms and financial returns arose from an analysis of the General Inquirer dictionary - a well-used dictionary in the literature on sentiment analysis. In addition to containing the categories of negative and positive evaluation, this dictionary also contains the ethically evaluative categories of vice and virtue. Intuitively the ethically evaluative categories comprise

a subset of the more general evaluative categories, and terms that are marked as vice or virtue in the GI dictionary are also labelled as negative or positive respectively. This led me to wonder whether vice terms could be used to derive a proxy for investor sentiment, in a similar manner to the way in which negative terms have been used in the past. As existing research has observed a strong relationship between negative affect in text and asset returns in financial markets, I sought to investigate whether it was really a subset of these terms - the vice terms - that were really driving this effect. As a result, this thesis built a proxy of investor sentiment from the GI's vice terms, and compared the results to an equivalent proxy of investor sentiment, constructed using the GI's negative terms.

To this end I constructed a text analysis system, which was capable of constructing a sentiment time series from unstructured text. This system used the General Inquirer dictionary and its associated disambiguation rules to compute the frequencies of negative and vice terms in text. Specifically, it takes unstructured textual data as input and produces time series of negative and vice terms as output that can be used in statistical modelling. This system is capable of working with unstructured textual data at any temporal frequency (daily, weekly, monthly etc.). Additionally the system is capable of working with market data at different levels of market granularity (firm-level, industry-level, market-level). In essence the system is a dictionary approach, which involves calculating the frequency of affect and ethical terms in a given text. These affect terms are subsequently used in statistical models to examine their impact on financial returns.

The reader might be concerned that the sentiment variable is actually proxying for factual information rather than the general sentiment in the market. In this case, investors could be responding to factual events, and the observed relationship between negative/vice terms and financial returns could actually be measuring the relationship between factual information and returns. While this is a possibility, it is unlikely for several reasons. Firstly, the news sources used in this thesis are unlikely to contain any new information that investors do not already know. Most institutional investors have access to newswire services at their desks, and receive breaking news as it occurs. Any new information in the news would be quickly incorporated into changes in share prices. Conversely, throughout this thesis I examine the relationship between sentiment on day t and returns on day $t + 1$ - with a whole day's gap in between the two. Another reason

why the sentiment proxy is unlikely to be capturing factual information is because news would cause different reactions in different investors. While some investors might maintain a ‘long’ position in a security (in the belief that it would rise in value), others will maintain a ‘short’ position (in the belief that it will fall), and it is these contrary beliefs that keep a security’s price in equilibrium. A given news article is therefore likely to invoke different opinions for different investors - good for some investors, and bad for others, but ultimately balancing one another out. However, in this thesis I have shown that an increase in negative/vice terms leads to a decrease in next day’s returns, implying that the overall aggregate opinion in the market place is negative amongst all investors. However, one should not necessarily rule out an informational impact, and future work should perhaps attempt to address this issue, such as by incorporating other quantitative variables into the VAR models.

5.1 Main Findings

In this thesis I conducted three independent case studies comprising market-level, firm-level and industry-level market data. Each case study covered a distinct time period with its own set of business cycles and macroeconomic structure. Additionally, each study has its own textual corpus and so there was also variability between the sources used. In this section I discuss the major findings in each study, together with the overall findings of this thesis.

Market-Level Study

In the market-level study I extended a major existing paper Tetlock (2007), to explore the impact of ethical and affect terms on the DJIA index’s returns. I first replicated the results of the existing study using negative terms from the GI dictionary. Overall, I observed similar results, finding that a unit increase in negative terms led to an average -4.4 basis point decrease in the subsequent day’s returns. I then extended the results to the vice terms, which did not yield a statistically significant effect for the overall period.

I then performed a series of rolling regressions to explore the time-varying impact of both the negative and vice terms. Using a rolling window of 130 days resulted in

3,558 vector autoregression (VAR) models computed for each of the variables. The first lagged negative sentiment variable was found to be significant in 259 of these VAR models (7.28%). The first lag of the vice variable was observed to be significant in 211 models (5.94%). This suggests that the impact of negative and ethical terms on financial returns is confined to particular periods. However, one of the most interesting results is the lack of overlap between any of the significant models. Of all 259 models in which the first negative lag is significant, none of these models overlap with the equivalent vice models. This suggests that the vice terms cannot be simply substituted for the negative terms, as the negative terms contain information about returns that is not captured in the vice terms alone. Additionally, this suggests that the vice terms contain information about returns that is lost when using the broader set of negative terms.

Firm-Level Study

In the firm-level study I explored the effect of negative and vice terms upon Enron's returns. I first noticed that there were several outliers in the Enron series that were affecting the fitted model. After winsorising these outliers to minimize their effect, I found that both the negative and the vice terms exhibited similar effects on the returns. A unit increase in the first lag of the negative variable led to an average -10.1 basis point decrease in the following day's returns. A unit increase in the equivalent lag's vice variable led to a -13.4 decrease in the following day's returns. Both the negative and vice variables were also found to be jointly significant, however the negative terms yielded a slightly stronger effect.

In the firm-level study the relationship between sentiment and returns was found to be bi-directional, in the sense that while sentiment had a significant impact on next day returns, returns were also found to have a significant impact on sentiment. Specifically, an increase in returns was observed to have a statistically significant decrease in the number of negative and vice terms and vice versa. This is suggestive of a feedback phenomenon, whereby sentiment impacts returns, which in turn impact sentiment.

As in the market-level study I also performed a series of rolling regression models, using the same rolling parameters to ensure comparability. This resulted in 1,062 VAR models for the negative and vice variables. Of these models, the first lag of the negative variable was found to be statistically significant in 52 (4.90%). Equivalently,

the first lag of the vice variable was found to be statistically significant in 147 of these models (13.84%). Interestingly, there was only a small overlap between the statistically significant negative models and the statistically significant vice models, suggesting as in the first study that the vice terms cannot be simply substituted for the negative terms.

Industry-Level Study

In the industry-level study I explored the effect of ethical and affect terms upon the NYSE's ARCA pharmaceutical index. To the best of my knowledge, this is the first research to look at this particular index. The main novelty of this particular study was the incorporation of the domain override dictionary. Prior to incorporating the domain override dictionary, none of the variables in either the negative or the vice models were found to be significant either individually or jointly. However, it transpired that a lot of the terms which the GI considered negative (vice) did not have ethical or evaluative connotations in this particular domain. Obvious examples were words such as 'cancer', 'symptom' and 'disease', which accounted for the most frequent evaluative terms in the corpus. Thus the main finding from this study was that the General Inquirer sentiment dictionary needs to be adapted for use in the pharmaceutical domain.

After compiling a domain-specific lexicon of pharmaceutical terms, it was observed that during expansionary periods a unit increase in the prior day's vice terms led to an average -5.7 basis point impact on the following day's returns. No significantly equivalent effect was observed in the expansionary period for the negative terms. Both the vice and negative terms also impacted upon returns during the recessionary period, however the results were observed to be quite counter-intuitive: a unit increase in negative terms at lag $t - 4$ led to an average 14.3 basis point increase in returns at time t . Conversely, a unit increase in vice terms at time $t - 2$ had a -16.7 basis point impact on returns at time t . The reason for these counter-intuitive results is probably due to the changing variance in the recessionary vs. expansionary periods. During the recessionary period, the NYSE ARCA returns exhibited a standard deviation of 1.80×10^2 . Conversely, during the expansionary period, the returns exhibited a standard deviation of only 0.89×10^2 - just under half the variance. The VAR model assumes that the variance remains constant over time, however as Figure 4.5 clearly shows there is a large amount of volatility clustering during this particular period. A possible remedial

solution for future research, would be to model this time-varying volatility. This could be achieved, for example, using the ARCH family of models which would model the time-varying heteroscedasticity in the returns.

A series of rolling regression estimations were also computed for the industry-level study. In total 2,071 models were fitted, of which the first lag of the negative variable was significant in 93 (4.49%). Conversely, the first lag of the vice variable was significant in 98 (4.73%). The findings thus support those in the previous study, namely that the impact of negative and vice terms varies over time. Of the periods in which the negative and vice models exhibit a significant impact on returns, only 33 of these periods overlap. This supports the observation in the previous 2 case studies, namely that the vice terms cannot be simply substituted for the negative terms.

Overall Findings

The main overall finding from this research is that the impact of both the negative and vice terms returns is confined to particular time periods. This is found to be the case across all three layers of market granularity, and is consistent with the previous literature in Ahmad et al. (2016) and García (2013). Overall this supports the idea of a dynamic marketplace, in which the macroeconomic structure is constantly evolving (this supports the adaptive market hypothesis (Lo (2004)) which states that financial markets evolve over time). This leads to the conclusion that the markets are *sometimes* inefficient, and during these times the news impacts returns.

At the outset of this report I established that there is a large overlap in the GI dictionary between words tagged as negative and the words tagged as vice. This led to the question of whether the vice terms had a similar impact to the negative terms upon financial returns. By and large the coefficients of the vice variables exhibit similar signs in the coefficients to those of the negative variables. However, the magnitude and significance of these coefficients changes according to the textual data and time period of study. The rolling regression results suggest that while the vice terms impact on financial returns during certain periods, the negative impact on financial returns during others. There is relatively little overlap between these periods, suggesting that each variable is capturing information about the returns that cannot be captured using only one variable in isolation.

In both the firm-level and industry-level studies I also conducted regression tests using the negative and vice terms as the dependent variable. Here the idea was to see whether financial returns contained any useful information about the sentiment measures. While the relationship between ethical/evaluative and financial returns was observed to be unidirectional in the industry-level study (historical ethical/evaluative terms impacted returns but not vice versa), the relationship was observed to be bidirectional in the firm-level study. That is, while an increase (decrease) in historical vice and negative terms led to a decrease (increase) in future financial returns, it was also the case that an increase (decrease) in financial returns led to a decrease (increase) in the number of negative and vice terms. Thus in the firm-level study there appeared to be some sort of feedback phenomenon whereby sentiment influenced returns which in turn influenced sentiment. This bi-directional finding is consistent with the results in Tetlock (2007), Tetlock et al. (2008) and García (2013). The lack of bi-directionality in the industry-level study is not a major cause for concern. It may be the case that changes in the pharmaceutical index are simply less newsworthy than changes in firm-level or market-level returns. Alternatively, it might be the case that investors hold longer positions in the pharmaceutical industry (perhaps as a result of the long research and development process), and therefore are unperturbed by short term deviations in market value. A third speculation could be that investors are simply used to changes in this index, which exhibited higher volatility than the market-level or firm-level returns. It would be interesting to see whether a bi-directional effect exists using firm-level pharmaceutical returns.

5.2 Limitations & Future Work

In this thesis I adopted a uniform weighting scheme, whereby evaluative words and phrases have an equal intensity. In reality, words can have different semantic strengths, with some words have a stronger force than others⁵⁷. As a result, one possible avenue for future work would be to explore the impact of different (non-binary) weighting schemes, which give larger weights to words with a stronger sentiment intensity. One idea could be to weight the terms according to their different level of impact upon

⁵⁷For instance *disastrous* has a stronger force than *bad*, despite both words having a negative polarity in the GI dictionary.

financial markets. Using a non-binary weighting scheme would also open the avenue to exploring the effects of other contextual valence shifters, such as intensifiers, hedges or diminishers which cannot be easily investigated using a binary weighting scheme.

One of the limitations with the rolling regression methodology is the issue of multiple comparisons. While computing a single test runs only a small risk of reporting a Type 1 (false positive) statistical error, comparing hypotheses computed from multiple regression tests simultaneously substantially increases this probability. It is often recommended to apply a Bonferroni correction (or other equivalent correction) in order to compensate for this risk. However, these tests are usually extremely conservative and could substantially increase the probability of generating Type 2 errors (covering up genuine effects that are present). This is particularly likely in the rolling window methodology where the data in each rolling window overlaps with the previous window. In this thesis I did not perform Bonferroni correction in the rolling window regressions. As a result, the reader should bear in mind that the reported results could possibly contain some false positives. Future work should therefore attempt to address this issue of multiple hypothesis testing, which has yet to receive substantive treatment in the rolling regression literature.

Another promising area for future work is to look at different levels of abstraction of the texts. In this thesis I analysed sentiment at the entire document level, but it may be the case that sentiment is confined to specific text portions or that the document level contains too much noise. For instance, in news articles the most important facts tend to appear in the headline and leading paragraphs of the article, with the latter paragraphs reserved for supplementary information. Consequently, it would be interesting to explore only these leading sentences, which would also significantly speed up the processing time of each text. It would also be interesting to explore how evaluative language modifies individual named entities. Rather than analysing the entire document, one could extract only sentences (or phrases) containing a particular entity, and analyse the sentiment at this level. This approach may require using some form of co-reference resolution in order to identify particular references to the target entities of interest.

The main goal of this thesis was to investigate the role of ethical terms in determining asset returns. In doing so, I used the vice terms from the GI dictionary, which comprised a subset of the negative terms traditionally used in sentiment analysis re-

search. While the coefficients of the two term-sets were generally similar in sign and magnitude, the rolling regressions revealed that there were periods in which the negative terms contained information about returns that was not found using only the vice terms in isolation. This begs the question of what predictive information is available in the entire set of negative terms, above and beyond that explained by only the subset of vice terms. One way of addressing this in future research would be to run additional regression models using the set of non-intersecting terms. This set would include the set of negative terms remaining, after all the vice terms have been extracted. If affect terms are grounded in ethical judgements, I would expect this new subset to yield little predictive power.

The industry-level case study in Chapter 4 made use of a domain override dictionary - essentially a list of domain-specific stop words and phrases that the system was instructed to ignore. While the override dictionary used in this study was a manually compiled list of words and phrases, it may be possible to compile this list automatically. For instance, if certain words and phrases occurred in the target corpus with a higher frequency than would be expected in a corpus of general English (such as the British National Corpus), it may be the case that these terms have some sort of specialist meaning in the target corpus beyond their everyday English usage. Future work could therefore attempt to automatically compile the domain override dictionary, using statistical keyword extraction techniques (such as a chi-square test or log-likelihood ratio) to identify words and phrases that occur with above-average frequencies.

In terms of the econometrics analysis, there are also a wide variety of opportunities for future research. In this thesis I have focussed mainly on vector autoregression models in order to make our analysis comparable to existing research. However, there are many other econometric models which could be explored in future research (such as event-based studies, panel-based studies, neural network models etc.). This thesis has also raised several unanswered questions, which present possible future areas for investigation. For instance what effect (if any) does weekend news have, and how can this be incorporated when markets are typically closed at weekends? Another question is what to do about news that is published in different time zones, or news published after the markets were closed. In this thesis many of the news sources did not have a precise time stamp and so no account was made for alternate time zones or market closures. Another question raised by this research was what to do about

joint interaction effects. I have seen that both sentiment and ethical terms have some impact upon financial markets, but it may be the case that the effect is interacted with other market variables. For instance, sentiment may have a greater impact when the market is more volatile, and therefore people may pay more attention to the news during turbulent market conditions.

Naturally, it would be interesting to look at different ethical terminology other than the terms from the GI dictionary. The results in this thesis are based upon ethical terms from the GI dictionary, but may not stand up to scrutiny when using alternative ethical dictionaries. One possibility would be to devise a statistical process for inducing a dictionary automatically such as by using keyword extraction, collocation analysis, or graph-based algorithms for determining synonymy (possibly in conjunction with an ontological structure such as WordNet). However, such a method would be difficult to evaluate without human intervention.

Finally, it would be interesting to see whether the disambiguation rules from the General Inquirer system could be derived automatically using some sort of machine learning procedure. This could potentially be used in place of (or in conjunction with) the domain ‘override’ dictionary to learn a set of contextual disambiguation rules for a given domain.

Bibliography

- Ahmad, K., Han, J., Hutson, E., Kearney, C. & Liu, S. (2016), ‘Media-expressed Negative Tone and Firm-level Stock Returns’, *Journal of Corporate Finance* **37**, 152–172.
- Aldridge, I. (2013), *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, 2nd edn, John Wiley & Sons Ltd, Hoboken, New Jersey.
- Antweiler, W. & Frank, M. Z. (2004), ‘Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards’, *The Journal of Finance* **59**(3), 1259–1294.
- Baker, M. & Wurgler, J. (2006), ‘Investor Sentiment and the Cross-Section of Stock Returns’, *The Journal of Finance* **61**(4), 1645–1680.
- Baker, R. E., Lembke, V. C., King, T. E. & Matulich, S. (1993), *Advanced Financial Accounting*, McGraw-Hill New York.
- Barber, B. M. & Odean, T. (2001), ‘Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment’, *Quarterly journal of Economics* pp. 261–292.
- Barber, B. M. & Odean, T. (2011), ‘The Behavior of Individual Investors’, *Social Science Research Network Working Paper Series* .
- Biber, D. & Finegan, E. (1989), ‘Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect’, *Text - Interdisciplinary Journal for the Study of Discourse* **9**(1), 93–124.
- Black, F. (1986), ‘Noise’, *Journal of Finance* **41**(3), 529–543.
- Blackford, J. U., Salomon, R. M. & Waller, N. G. (2009), ‘Detecting change in biological

- rhythms: A multivariate permutation test approach to fourier-transformed data’, *Chronobiology international* **26**(2), 258–281.
- Bollen, J., Maoa, H. & Zeng, X.-J. (2011), ‘Twitter Mood Predicts the Stock Market’, *Journal of Computational Science* **2**(1), 1–8.
- Boudoukh, J., Feldman, R., Kogan, S. & Richardson, M. (2013), Which News Moves Stock Prices? A Textual Analysis, Technical report, National Bureau of Economic Research.
- Brooks, C. (2014), *Introductory Econometrics for Finance*, 3rd edn, Cambridge University Press.
- Brown, G. W. & Cliff, M. T. (2004), ‘Investor Sentiment and the Near-term Stock Market’, *Journal of Empirical Finance* **11**(1), 1–27.
- Campbell, J. Y., Grossman, S. J. & Wang, J. (1993), ‘Trading Volume and Serial Correlation in Stock Returns’, *The Quarterly Journal of Economics* **108**(4), 905–939.
- Chan, W. S. (2003), ‘Stock Price Reaction to News and No-news: Drift and Reversal after Headlines’, *Journal of Financial Economics* **70**(2), 223–260.
- Chen, H., De, P., Hu, Y. J. & Hwang, B.-H. (2014), ‘Wisdom of crowds: The value of stock opinions transmitted through social media’, *Review of Financial Studies* **27**(5), 1367–1403.
- Cont, R. (2001), ‘Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues’, *Quantitative Finance* **1**, 223–236.
- Cook, J. A. & Ahmad, K. (2015), Behaviour and Markets: The Interaction Between Sentiment Analysis and Ethical Values?, in ‘International Conference on Intelligent Data Engineering and Automated Learning’, Springer, pp. 551–558.
- Cook, J. A., Zhao, Z. & Ahmad, K. (2016), Stylized Facts of Linguistic Corpora: Exploring the Lexical Properties of Affect in News, in ‘International Conference on Intelligent Data Engineering and Automated Learning’, Springer, pp. 494–502.
- Crystal, D. (2010), *The Cambridge Encyclopedia of Language*, 3rd edn, Cambridge University Press, Cambridge, UK.

- Damodar, G. (1999), *Basic Econometrics*, 3rd edn, McGraw-Hill Companies.
- Daniel, K., Hirshleifer, D. & Subrahmanyam, A. (1998), ‘Investor Psychology and Security Market Under and Overreactions’, *The Journal of Finance* **53**(6), 1839–1885.
- Das, S. R. (2014), ‘Text and Context: Language Analytics in Finance’, *Foundations and Trends in Finance* **8**(3), 145–261.
- Das, S. R. & Chen, M. Y. (2007), ‘Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web’, *Management Science* **53**(9), 1375–1388.
- Davis, A. K., Ge, W., Matsumoto, D. & Zhang, J. L. (2015), ‘The Effect of Manager-specific Optimism on the Tone of Earnings Conference Calls’, *Review of Accounting Studies* **20**, 639–673.
- Davis, A. K., Piger, J. M. & Sedor, L. M. (2012), ‘Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language’, *Contemporary Accounting Research* **29**(3), 845–868.
- Davis, A. K. & Tama-Sweet, I. (2012), ‘Managers’ Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A’, *Contemporary Accounting Research* **29**(3), 804–837.
- De Long, J. B., Shleifer, A., Summers, L. H. & Waldmann, R. J. (1990), ‘Noise Trader Risk in Financial Markets’, *Journal of Political Economy* **98**(4), 703–738.
- DeBondt, W. F. & Thaler, R. H. (1985), ‘Does the Stock Market Overreact’, *Journal of Finance*. **40**, 793–805.
- DellaVigna, S. & Pollet, J. M. (2009), ‘Investor Inattention and Friday Earnings Announcements’, *The Journal of Finance* **64**(2), 709–749.
- Demers, E. A. & Vega, C. (2010), ‘Soft Information in Earnings Announcements: News or Noise?’.
- URL:** <https://www.federalreserve.gov/pubs/ifdp/2008/951/ifdp951.pdf>
- Demers, E. A. & Vega, C. (2011), ‘Linguistic tone in earnings announcements: News or noise?’, *SSRN Working Paper Series* .

- Dickey, D. A. & Fuller, W. A. (1979), 'Distribution of the Estimators for Autoregressive Time Series With a Unit Root', *Journal of the American statistical association* **74**(366a), 427–431.
- Diction Software (2017), 'Diction Software', <http://www.dictionsoftware.com>. Online; Last Accessed 21 August 2017.
- Doran, J. S., Peterson, D. R. & Price, S. M. (2012), 'Earnings Conference Call Content and Stock Price: the case of REITs', *The Journal of Real Estate Finance and Economics* **45**(2), 402–434.
- Dunphy, D. C., Bullard, C. G. & Crossing, E. E. (1974), *Validation of the General Inquirer Harvard IV Dictionary*, Pisa Conference on Content Analysis.
- Engelberg, J. (2008), 'Costly Information Processing: Evidence from Earnings Announcements', *Social Science Research Network Working Paper Series*.
- Engelberg, J. E. & Parsons, C. A. (2011), 'The causal impact of media in financial markets', *The Journal of Finance* **66**(1), 67–97.
- Fama, E. (1970), 'Efficient Capital Markets: A Review of Theory and Empirical Work', *Journal of Finance* **25**(2), 383–417.
- Feldman, R., Govindaraj, S., Livnat, J. & Segal, B. (2010), 'Management's Tone Change, Post Earnings Announcement Drift and Accruals', *Review of Accounting Studies* **15**(4), 915–953.
- Ferguson, N. J., Philip, D., Lam, H. Y. T. & Guo, J. M. (2015), 'Media Content and Stock Returns: The Predictive Power of Press', *Multinational Finance Journal* **19**(1), 1–31.
- Ferris, S. P., Hao, G. Q. & Liao, S. M.-Y. (2013), 'The Effect of Issuer Conservatism on IPO Pricing and Performance', *Review of Finance* **17**, 993–1027.
- Fortune Magazine (2015), 'This 62-year-old drug just got 5,000% more expensive', <http://fortune.com/2015/09/21/turing-pharmaceuticals-drug-prices-daraprim/>. Online; Last Accessed 04 September 2016.
- Francis, W. N. & Kucera, H. (1967), 'The Brown Corpus: A Standard Corpus of Present-Day Edited American English'. Brown University Linguistics Department.

- Fung, G. P. C., Yu, J. X. & Lam, W. (2003), Stock Prediction: Integrating Text Mining Approach Using Real-time News, *in* ‘Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on’, IEEE, pp. 395–402.
- Gamon, M., Yano, T., Song, X., Apacible, J. & Pantel, P. (2013), Understanding Document Aboutness Step One: Identifying Salient Entities, Technical report, Technical Report MSR-TR-2013-73, Microsoft Research.
- García, D. (2013), ‘Sentiment During Recessions’, *The Journal of Finance* **68**(3), 1267–1300.
- Gilbert, E. & Karahalios, K. (2010), Widespread Worry and the Stock Market, *in* ‘ICWSM’, pp. 59–65.
- Gonzalez, A. (2011), Are You Still Trading Without News?, *in* G. Mitra & L. Mitra, eds, ‘The Handbook of News Analytics in Finance’, John Wiley & Sons Ltd., Chichester, West Sussex, chapter 15, pp. 311–314.
- Granger, C. W. J. (1969), ‘Investigating Causal Relations by Econometric Models and Cross-spectral Methods’, *Econometrica: Journal of the Econometric Society* pp. 424–438.
- Greene, W. H. (2000), *Econometric analysis*, Prentice Hall., New Jersey, USA.
- Groß-Klußmann, A. & Hautsch, N. (2011), ‘When Machines Read the News: Using Automated Text Analytics to Quantify High Frequency News-implied Market Reactions’, *Journal of Empirical Finance* **18**(2), 321–340.
- Hanley, K. W. & Hoberg, G. (2010), ‘The Information Content of IPO Prospectuses’, *Review of Financial Studies* **23**(7), 2821–2864.
- Henry, E. (2006), ‘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.
- Henry, E. (2008), ‘Are Investors Influenced by How Earnings Press Releases are Written?’, *Journal of Business Communication* **45**(4), 363–407.

- Henry, E. & Leone, A. J. (2015), ‘Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone’, *The Accounting Review* **91**(1), 153–178.
- Heston, S. L. & Sinha, Nitish, R. (2016), ‘News versus Sentiment: Predicting Stock Returns from News Stories’, *Finance and Economics Discussion Series 2016-048*.
- Ho, K.-Y., Shi, Y. & Zhang, Z. (2013), ‘How Does News Sentiment Impact Asset Volatility? Evidence from Long Memory and Regime-switching Approaches’, *The North American Journal of Economics and Finance* **26**, 436–456.
- Hogenboom, A., Iterson, P. V., Heerschop, B., Frasincar, F. & Kaymak, U. (2011), Determining Negation Scope and Strength in Sentiment Analysis, in ‘Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on’, pp. 2589–2594.
- Huang, X., Teoh, S. H. & Zhang, Y. (2014), ‘Tone Management’, *The Accounting Review* **89**(3), 1083–1113.
- Hunston, S. (2010), *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*, Routledge.
- Investopedia (2005), ‘What is an earnings conference call?’, <http://www.investopedia.com/ask/answers/04/052104.asp>. Online; Last Accessed 21 August 2017.
- Jaffe, A. (2009), *Stance: Sociolinguistic Perspectives*, OUP USA.
- Jegadeesh, N. & Titman, S. (1993), ‘Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency’, *The Journal of finance* **48**(1), 65–91.
- Jegadeesh, N. & Wu, D. (2013), ‘Word Power: A New Approach for Content Analysis’, *Journal of Financial Economics* **110**, 712–729.
- Jin, F., Self, N., Saraf, P., Butler, P., Wang, W. & Ramakrishnan, N. (2013), Forex-foreteller: Currency Trend Modeling using News Articles, in ‘Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 1470–1473.
- Kahneman, D. & Tversky, A. (1979), ‘Prospect Theory: An Analysis of Decision under Risk’, *Econometrica* **47**(2), 263–292.

- Kahneman, D. & Tversky, A. (1972), ‘Subjective Probability: A Judgment of Representativeness’, *Cognitive psychology* **3**(3), 430–454.
- Kearney, C. & Liu, S. (2014), ‘Textual Sentiment in Finance: A Survey of Methods and Models’, *International Review of Financial Analysis* **33**, 171–185.
- Kelly, E. F. & Stone, P. (1975), *Computer Recognition of English Word Senses*, North-Holland Linguistic Series, North-Holland Publishing Company.
- Kennedy, A. & Inkpen, D. (2006), ‘Sentiment classification of movie reviews using contextual valence shifters’, *Computational Intelligence* **22**(2), 110–125.
- Keynes, J. M. (1936), *The General Theory of Employment, Interest and Money*, Macmillan, London, UK.
- Kibbe, W. A., Arze, C., Felix, V., Mitiraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D. et al. (2014), ‘Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge Through Disease Data’, *Nucleic acids research* .
- Kim, S.-H. & Kim, D. (2014), ‘Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns’, *Journal of Economic Behavior & Organization* **107**, 708–729.
- Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S. & Smith, N. A. (2009), Predicting Risk from Financial Reports with Regression, in ‘Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics’, NAACL ’09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 272–280.
- Kothari, S. P., Li, X. & Short, J. E. (2009), ‘The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis’, *The Accounting Review* **84**(5), 1639–1670.
- Lasswell, H. (1927), *Propaganda Techniques in the World War*, Knopf, New York.
- Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J. (2000), Mining of Concurrent Text and Time Series, in ‘In Proceedings of the 6th ACM SIGKDD

- International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining’.
- Leinweber, D. & Sisk, J. (2011), Relating News Analytics to Stock Returns, *in* G. Mitra & L. Mitra, eds, ‘The Handbook of News Analytics in Finance’, John Wiley & Sons Ltd, Chichester, West Sussex, UK, pp. 149–172.
- Li, F. (2006), ‘Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?’, *Social Science Research Network Working Paper Series* .
- Li, F. (2010a), ‘Textual Analysis of Corporate Disclosures: A Survey of the Literature’, *Journal of Accounting Literature* **29**, 143–165.
- Li, F. (2010b), ‘The Information Content of Forward-looking Statements in Corporate Filings - a Naïve Bayesian Machine Learning Approach’, *Journal of Accounting Research* **48**, 1049–1102.
- Li, X., Xie, H., Chen, L., Wang, J. & Deng, X. (2014), ‘News Impact on Stock Price Return Via Sentiment Analysis’, *Knowledge-Based Systems* **69**, 14–23.
- Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.
- Liu, B. (2015), *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*, Cambridge University Press, Cambridge, UK.
- Liu, B. & Zhang, L. (2012), A Survey of Opinion Mining and Sentiment Analysis, *in* C. C. Aggarwal & C. Zhai, eds, ‘Mining Text Data’, Springer US, Boston, MA, pp. 415–463.
- Lo, A. W. (2004), ‘The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective’, *Journal of Portfolio Management*, *Forthcoming* .
- Loughran, T. & McDonald, B. (2011), ‘When Is a Liability Not a Liability? Textual Analysis, Dictionaries and 10-Ks’, *The Journal of Finance* **66**(1), 35–65.
- Loughran, T. & McDonald, B. (2015), ‘The Use of Word Lists in Textual Analysis’, *Journal of Behavioral Finance* **16**(1), 1–11.
- Loughran, T. & McDonald, B. (2016), ‘Textual Analysis in Accounting and Finance: A Survey’, *Journal of Accounting Research* **54**(4), 1187–1230.

- Luss, R. & d'Aspremont, A. (2015), 'Predicting Abnormal Returns from News Using Text Classification', *Quantitative Finance* **15**(6), 999–1012.
- Makridakis, S., Wheelwright, S. C. & Hyndman, R. J. (1998), *Forecasting: Methods and Applications*, 3rd edn, John Wiley & Sons, New York, USA.
- Mandelbrot, B. B. (1963), 'The Variation of Certain Speculative Prices', *The Journal of Business* **36**(4), 394–419.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014), The Stanford CoreNLP Natural Language Processing Toolkit, in 'Association for Computational Linguistics (ACL) System Demonstrations', pp. 55–60.
- Mao, H., Counts, S. & Bollen, J. (2015), Quantifying the Effects of Online Bullishness on International Financial Markets, Technical report, European Central Bank.
- Martin, J. R. & White, P. R. (2003), *The Language of Evaluation*, Palgrave Macmillan, Basingstoke, UK.
- MathWorks (2014), 'Rolling-Window Analysis of Time-Series Models', <https://uk.mathworks.com/help/econ/rolling-window-estimation-of-state-space-models.html>. Online; Last Accessed 28 November 2016.
- Mitra, L. & Mitra, G. (2011), Applications of News Analytics in Finance: A Review, in G. Mitra & L. Mitra, eds, 'The Handbook of News Analytics in Finance', John Wiley & Sons Ltd., Chichester, West Sussex, chapter 1, pp. 1–38.
- Mittermayer, M.-A. (2004), Forecasting Intraday Stock Price Trend with Text Mining Techniques, in 'Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICS), IEEE Computer Society'.
- Munz, M. (2011), Measuring the Value of Media Sentiment: a Pragmatic Review, in G. Mitra & L. Mitra, eds, 'The Handbook of News Analytics in Finance', John Wiley & Sons Ltd., Chichester, West Sussex, chapter 4, pp. 109 – 128.
- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2014), 'Text Mining for Market Prediction: A Systematic Review', *Expert Systems with Applications* **41**(16), 7653–7670.

- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2015), ‘Text Mining of News-headlines for FOREX Market Prediction: A Multi-Layer Dimension Reduction Algorithm with Semantics and Sentiment’, *Expert Systems with Applications* **42**, 306–324.
- New York Stock Exchange (2014), ‘The New York Stock Exchange ARCA Pharmaceutical Index (DRG)’, https://www.nyse.com/publicdocs/nyse/indices/nyse_arca_pharmaceutical_index.pdf. Online; Last Accessed 04 September 2016.
- Nickerson, R. S. (1998), ‘Confirmation Bias: A Ubiquitous Phenomenon in Many Guises’, *Review of general psychology* **2**(2), 175.
- Niederhoffer, V. (1971), ‘The Analysis of World Events and Stock Prices’, *The Journal of Business* **44**(2), 193–219.
- Oh, C. & Sheng, O. (2011), ‘Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement’.
- Oliveira, N., Cortez, P. & Areal, N. (2013), On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume, in L. Correia, L. P. Reis & J. Cascalho, eds, ‘Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, September 9-12, 2013. Proceedings’, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 355–365.
- Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., Danila, M. I., Feng, G. & Chisholm, R. L. (2009), ‘Annotating the Human Genome with Disease Ontology’, *BMC genomics* **10**(Suppl 1), S6.
- Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. (1957), *The Measurement of Meaning*, University of Illinois Press, Urbana, Illinois.
- Pang, B., Lee, L. & Vaithyanathan, S. (2002), Thumbs up? Sentiment Classification Using Machine Learning Techniques, in ‘Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10’, Association for Computational Linguistics, pp. 79–86.

- Paranjpe, D. (2009), Learning document aboutness from implicit user feedback and document structure, *in* ‘Proceedings of the 18th ACM Conference on Information and Knowledge Management’, ACM, pp. 365–374.
- Polanyi, L. & Zaenen, A. (2006), *Contextual Valence Shifters*, Springer Netherlands, Dordrecht, pp. 1–10.
- Pompian, M. (2012), ‘Behavioral Finance and Investor Types’, *Private Wealth Management Feature Articles* **2012**(1), 1–3.
- Price, S. M., Doran, J. S., Peterson, D. R. & Bliss, B. A. (2012), ‘Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone’, *Journal of Banking & Finance* **36**(4), 992–1011.
- Rogers, J. L., Buskirk, A. V. & Zechman, S. L. C. (2011), ‘Disclosure Tone and Shareholder Litigation’, *The Accounting Review* **86**(6), 2155–2183s.
- Ruppert, D. (2004), Trimming and winsorization, *in* S. Kotz, N. Balakrishnan, C. Read & B. Vidakovic, eds, ‘Encyclopedia of Statistical Sciences’, 2 edn, Vol. 14, John Wiley & Sons, Inc.
- Sadique, S., In, F. H. & Veeraraghavan, M. (2008), ‘The Impact of Spin and Tone on Stock Returns and Volatility: Evidence from Firm-Issued Earnings Announcements and the Related Press Coverage’, *Social Science Research Network Working Paper Series* .
- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G. & Kibbe, W. A. (2012), ‘Disease Ontology: a Backbone for Disease Semantic Integration’, *Nucleic acids research* **40**(D1), D940–D946.
- Schumaker, R. P. & Chen, H. (2006), Textual Analysis of Stock Market Prediction Using Financial News Articles, *in* ‘In 12th American Conference on Information Systems (AMCIS)’, IEEE Computer Society, Los Alamitos, CA.
- Schumaker, R. P. & Chen, H. (2009), ‘A Quantitative Stock Prediction System Based on Financial News’, *Information Processing & Management* **45**(5), 571–583.
- Shiller, R. J. (2000), *Irrational Exuberance*, Princeton University Press, Princeton, NJ.

- Shleifer, A. (2000), *Inefficient Markets: An Introduction to Behavioural Finance*, OUP Oxford.
- Shynkevich, Y., McGinnity, T., Coleman, S. A. & Belatreche, A. (2016), ‘Forecasting Movements of Health-care Stock Prices Based on Different Categories of News Articles using Multiple Kernel Learning’, *Decision Support Systems* **85**, 74–83.
- Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. & Deng, X. (2013a), *Exploiting Topic Based Twitter Sentiment for Stock Prediction*, Vol. 2, Association for Computational Linguistics (ACL), pp. 24–29.
- Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. & Deng, X. (2013b), ‘Exploiting Topic based Twitter Sentiment for Stock Prediction’, *ACL (2)* **2013**, 24–29.
- Sims, C. A. (1980), ‘Macroeconomics and Reality’, *Econometrica* **48**, 1–48.
- Sims, R. R. & Brinkmann, J. (2003), ‘Enron ethics (or: culture matters more than codes)’, *Journal of Business ethics* **45**(3), 243–256.
- Sinha, N. R. (2016), ‘Underreaction to News in the US Stock Market’, *Quarterly Journal of Finance* **6**(2).
- Sprenger, T. O., Sandner, P. G., Tumasjan, A. & Welpe, I. M. (2014), ‘News or Noise? Using Twitter to Identify and Understand Company-specific News Flow’, *Journal of Business Finance & Accounting* **41**(7-8), 791–830.
- Stone, P. J. (1969), The Analysis of Communication Content, in G. Gerbner, O. Holsti, K. Krippendorff, W. Paisley & P. J. Stone, eds, ‘Improved Quality of Content Analysis Categories: Computerized Disambiguation Rules For High Frequency English Words’, Krieger, pp. 199–233.
- Stone, P. J., Dunphy, D. C. & Smith, M. S. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, MIT press, Massachusetts.
- Taylor, S. J. (2011), *Asset Price Dynamics, Volatility, and Prediction*, Princeton University Press.
- Temin, P. (2010), ‘The Great Recession & the Great Depression’, *Daedalus* **139**(4), 115–124.

- Tetlock, P. C. (2007), ‘Giving Content to Investor Sentiment: The Role of Media in the Stock Market’, *Journal of Finance* **62**(3), 1139–1168.
- Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. (2008), ‘More than Words: Quantifying Language to Measure Firms’ Fundamentals’, *The Journal of Finance* **63**(3), 1437–1467.
- Thaler, R. H., ed. (2005), *Advances in Behavioral Finance*, Vol. 2, Princeton University Press, Princeton, USA.
- Thomas, J. D. & Sycara, K. (2000), ‘Integrating Genetic Algorithms and Text Learning for Financial Prediction’, *Data Mining with Evolutionary Algorithms* pp. 72–75.
- Tumarkin, R. & Whitelaw, R. F. (2001), ‘News or Noise? Internet Postings and Stock Prices’, *Financial Analysts Journal* **57**(3), 41–51.
- Tversky, A. & Kahneman, D. (1975), Judgment Under Uncertainty: Heuristics and Biases, in ‘Utility, probability, and human decision making’, Springer, pp. 141–162.
- Tversky, A. & Kahneman, D. (1985), The Framing of Decisions and the Psychology of Choice, in ‘Environmental Impact Assessment, Technology Assessment, and Risk Analysis’, Springer, pp. 107–129.
- Twedt, B. & Rees, L. (2012), ‘Reading Between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts’ Reports’, *Journal of Accounting and Public Policy* **31**(1), 1–21.
- US Department of Justice (2012), ‘GlaxoSmithKline to Plead Guilty and Pay \$3 Billion to Resolve Fraud Allegations and Failure to Report Safety Data’, <https://www.justice.gov/opa/pr/glaxosmithkline-plead-guilty-and-pay-3-billion-resolve-fraud-allegations-and-failure-report>. Online; Last Accessed 04 September 2016.
- Vu, T.-T., Chang, S., Ha, Q. T. & Collier, N. (2012), An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter, in ‘Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data’, COLING, Mumbai, pp. 23–38.

- Wall Street Journal (2015), ‘Merck is Accused of Stonewalling Over Effectiveness of Mumps Vaccine’, <http://blogs.wsj.com/pharmalot/2015/06/08/merck-is-accused-of-stonewalling-over-effectiveness-of-mumps-vaccine/>. Online; Last Accessed 04 September 2016.
- Wisniewski, T. P. & Sina Yekini, L. (2015), ‘Stock Market Returns and the Content of Annual Report Narratives’, *Accounting Forum* **39**, 281–294.
- Wüthrich, B., Permuntilleke, D., Leung, S., Lam, W., Cho, V. & Zhang, J. (1998), ‘Daily Prediction of Major Stock Indices from Textual WWW Data’, *HKIE Transactions* **5**(3), 151–156.
- Yu, Y., Duan, W. & Cao, Q. (2013), ‘The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach’, *Decision Support Systems* **55**(4), 919–926.
- Zhang, J. L., Härdle, W. K., Chen, C. Y. & Bommers, E. (2016), ‘Distillation of News Flow Into Analysis of Stock Reactions’, *Journal of Business & Economic Statistics* **34**(4), 547–563.
- Zhang, W. & Skiena, S. (2010), Trading Strategies to Exploit Blog and News Sentiment, in ‘Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media’.
- Zhang, X., Fuehres, H. & Gloor, P. A. (2011), ‘Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”’, *Procedia - Social and Behavioral Sciences* **26**, 55–62.
- Zhang, X., Fuehres, H. & Gloor, P. A. (2012), Predicting Asset Value Through Twitter Buzz, in ‘Advances in Collective Intelligence 2011’, Springer, pp. 23–34.

Appendices

Appendix A

GI's Contextual Disambiguation Rules

In this thesis I utilised the General Inquirer's (GI's) contextual disambiguation rules to map polysemous words onto their appropriate sense. The basic structure of the GI's contextual disambiguation rule is given in Table A.1, with descriptions taken from Kelly & Stone (1975). An example set of disambiguation rules for the word 'kind' is given in A.2. Although there are many other subtleties to these contextual disambiguation rules, the examples in A.2 indicate a variety of polysemous problems which can be dealt with using the GI's disambiguation rules. It is perhaps important to note that the disambiguation rules work at the sentence level, and therefore any extra-sentential contextual information is disregarded. The authors who constructed the rules found that the majority of words could be disambiguated from a key-word in context (KWIC) approach, and thus deemed extra-sentential information as an unnecessary overhead.

Table A.1: Basic structure of the GI’s contextual disambiguation rules. Each of the contextual rules contains all the elements given in the test element column. The descriptions are taken from p14-15 of Kelly & Stone (1975)

Test Element	Description
Test type flag	Indicating whether the test is looking for specific words or specific tags
Range	Indicating the context window (where the test should begin or end in the sentence)
Search-type Flag	Indicating the arrangement required of the test items for the test to succeed. Possible options include: • Single test item (occurring anywhere in the range) • Multiple test items: – Any one anywhere in the range – All in the range and in text order – All in the range, in text order and adjacent – All in the range and all on one word (tag tests only)
Items list	Specifying the items to be searched for (set of words/tags). This can either be a list of words or a list of tags (semantic or syntactic).
Success and failure branches	To be executed depending on the outcome of the test. Possible options include: • Pass to the next rule (rules are ordered by most to least frequent senses) • Transfer to another rule • Transfer to another dictionary entry and execute those rules • Apply a specific sense number and any applicable tags; cease to execute further rules • Treat the specified context window as an idiomatic expression; apply any applicable tags and cease to execute further rules.

Table A.2: Example context rules for the word ‘kind’. This snapshot of rules is taken directly from the GI System’s disambiguation rule list. The first two arguments (e.g. K+1, K+2) indicate the start and end positions of the context window to be searched. Here, ‘K’ represents the position of the word in the sentence that is currently being disambiguated. The third and fourth arguments represent the success and failure branches respectively. The fifth argument represents a dot delimited list of items (word or tags) to search for within the context window specified in the first two arguments. Finally, the ‘T’ and ‘W’ at the start of the function names indicate whether the test is a ‘tag’ or ‘word’ test respectively, the ‘OR’ indicates logical OR (any items will suffice), and the ‘K’ in ‘TORK’ indicates that the keyword may intervene. Other possible test options also exist but these are excluded here from these examples. The interested reader is invited to consider Kelly & Stone (1975) for further examples.

Test Rule	Description
TOR(K+0,K+0,,11,LY.	Look for the presence of the LY tag on the current word. If the tag is present, then proceed to the next rule in the sequence, otherwise transfer directly to rule 11. Note that the LY tag would usually only apply to the word ‘kindly’ in the text; the stemming algorithm would remove the -ly suffix, and assign the LY marker tag.
TORK(K-1,K-2,,15,SUPV.PUNC.DEF1.	Look for the presence of any of the tags SUPV, PUNC or DEF1 on the proceeding two words in the sentence. If any of these tags occur on any of the two proceeding words, then proceed to the next rule. Otherwise, transfer directly to rule 15.
WOR(K+1,K+1,DELID(3),,OF.	Check if the next word in the sentence is the word ‘of’. If so, apply sense 3 to the word ‘kind’ and treat the sequence ‘kind of’ as an idiom. If not, transfer to the next rule in the sequence
WOR(K-1,K-1,DELID(4),,IN.	Check if the previous word in the sentence is the word ‘in’. If so, apply sense 4 to the word ‘kind’ and treat the sequence ‘in kind’ as an idiom. If not, transfer to the next rule in the sequence.
TOR(K-1,K-3,APLY(2),APLY(1),BE.	Check for the present of the tag ‘BE’ on the previous two words in the sentence. If the tag is present, then apply sense 2. Otherwise, apply sense 1.

Appendix B

Stemming Algorithm

In this thesis I used the stemming algorithm designed by Kelly and Stone (Kelly & Stone (1975), pp. 95-97), which was originally created in conjunction with the GI's disambiguation procedures. The algorithm is applied to individual words during the lookup phase. The stemming algorithm is implemented as a recursive solution, which performs a dictionary lookup at each stage. If a word is found in the GI (or an override) dictionary, then execution of the algorithm is terminated. If not, then the algorithm continues to stem the word until either a lookup is found or else no further stemming procedures are possible. During each iteration of the stemming algorithm a set of suffix tags may be applied. For instance, if the suffix -s is removed from a word, then the tag 'S' is applied. At this point, following any further unsuccessful lookups, the program can re-call the stemming function but transferring to another part of the stemming routine. The complete stemming code, adapted from Kelly & Stone (1975), is shown below:

Algorithm 1 Stemming algorithm

```
1: procedure STEM(word) ▷ Word object
2:   if word ends with “-S” then go to 11
3:   else if word ends with “-D” then go to 21
4:   else if word ends with “-G” then go to 29
5:   else if word ends with “-R” then go to 40
6:   else if word ends with “-Y” then go to 48
7:   else if word ends with “-T” then go to 56
8:   end if
9:   return ▷ Stop stemming
10:
11:   S:
12:   if len(word) ≥ 2 && !word.endsWith(“-OUS”) && !word.endsWith(“-SS”)
      then
13:     remove “-S” ending;
14:     add S marker tag;
15:     if lookup() == FALSE then ▷ Perform lookup in GI and override
      dictionary
16:       go to 63
17:     end if
18:   end if
19:   return ▷ Stop stemming
20:
21:   D:
22:   if word ends with “-ED” then
23:     if word does not have LY marker tag then ▷ (e.g. ‘DECIDEDLY58’)
24:       add ED marker tag.
25:     end if
26:     go to 76
27:   end if
28:   return ▷ Stop stemming
```

```

29:  G:
30:  if len(word) ≥ 3 && word.endsWith("ING") then
31:      remove "-ING" suffix
32:      add "-E" suffix
33:      if LY suffix tag not already assigned then
34:          assign ING suffix tag
35:      end if
36:      if lookup() == FALSE then go to 91 end if ▷ Perform lookup in GI and
      override dictionary
37:  end if
38:  return ▷ Stop stemming
39:
40:  R:
41:  if len(word) ≥ 2 && word.endsWith("ER") then
42:      remove "-R"
43:      assign ER suffix tag
44:      if lookup() == FALSE then go to 76 end if ▷ Perform lookup in GI and
      override dictionary
45:  end if
46:  return ▷ Stop stemming
47:
48:  Y:
49:  if len(word) ≥ 2 && word.endsWith("LY") then
50:      remove "-LY"
51:      add LY suffix tag
52:      if lookup() == FALSE then go to 99 end if ▷ Perform lookup in GI and
      override dictionary
53:  end if
54:  return ▷ Stop stemming
55:
56:  T:
57:  if len(word) ≥ 3 && !word.endsWith("EST") then
58:      remove "-ST" suffix
59:      add EST suffix tag
60:      if lookup() == FALSE then go to 76 end if ▷ Perform lookup in GI and
      override dictionary
61:  end if
62:  return ▷ Stop stemming

```

```

63:  ES:
64:  if len(word)  $\geq$  2 then
65:      if word.endsWith("ING") then
66:          go to 29
67:      else if word.endsWith("E") then
68:          remove 'E' suffix
69:          if word.endsWith("I") then
70:              change 'I' suffix to 'Y'
71:          end if
72:      end if
73:  end if
74:  return                                ▷ Stop stemming
75:
76:  ERDT:
77:  remove 'E' suffix
78:  if lookup() == FALSE then go to 81 end if    ▷ Perform lookup in GI and
override dictionary
79:  return                                ▷ Stop stemming
80:
81:  ERDT2:
82:  if len(word)  $\geq$  2 then
83:      if word.endsWith("I") then
84:          change last letter to 'Y'
85:      else if word has double last letter then
86:          remove last letter
87:      end if
88:  end if
89:  return                                ▷ Stop stemming
90:
91:  G2:
92:  remove last letter                    ▷ The added 'E' from previous step
93:  if lookup() == FALSE then go to 96 end if    ▷ Perform lookup in GI and
override dictionary
94:  return                                ▷ Stop stemming
95:
96:  G3:
97:  if word has double last letter then remove last letter end if
98:  return                                ▷ Stop stemming

```

```

99:   Y2:
100:   if word.endsWith("B") or word.endsWith("P") or word.endsWith("T") then
101:       add "LE" suffix                                ▷ 'MISERABLY' → 'MISERABLE'
102:   else if word.endsWith('I') then
103:       change last letter to 'Y'
104:   else if word.endsWith('L') then
105:       if len(word) > 3 then
106:           if word.endsWith("CAL") then
107:               remove 'AL' suffix                        ▷ 'SPECIFICALLY' → 'SPECIFIC'
108:           else
109:               add an 'E' suffix                            ▷ 'WHOLLY' → 'WHOLE'
110:               if lookup() == FALSE then go to 119 end if ▷ Perform lookup in
               GI and override dictionary
111:           end if
112:       else if word.endsWith('U') then add 'E'            ▷ 'TRULY' → 'TRUE'
113:       else if word.endsWith('G') then go to 29
114:       else if word.endsWith('D') then go to 21
115:       end if
116:   end if
117:   return                                              ▷ Stop stemming
118:
119:   Y3:
120:   substitute 'L' for the added 'E'                    ▷ 'FULLY' → 'FULE' → 'FULL'
121:   return                                              ▷ Stop stemming
122: end procedure

```

Appendix C

Durbin-Watson Statistics

One of the assumptions of regression using time series data is that there is no autocorrelation in the residuals. To check for this, I computed Durbin-Watson statistics (Equation 3.7), which detects the presence of autocorrelation in the residuals. The test posits the following null and alternative hypotheses:

H_0 (null hypothesis): no autocorrelation in the residuals

H_1 (alternative hypothesis): autocorrelation is present in the residuals.

C.1 Enron Case Study

Table C.1 shows results of the Durbin-Watson tests for the model fitted in section 4.3.4.2. The test statistics show that there is insufficient evidence to reject the null hypothesis of no autocorrelation in the residuals. Thus, the fitted model meets the assumption of uncorrelated residuals. While the results here are for the fitted model using Enron's original pre-winsorised returns, similar results are observed for all of the other fitted models.

Table C.1: Results of Durbin-Watson test for autocorrelation in the residuals. The negative column denotes the VAR model fitted using the frequency of negative terms as a proxy of investor sentiment. The vice column denotes the equivalent model using the frequency of vice terms.

	Negative	Vice
Durbin-Watson	1.99	1.99
p -value	0.80	0.78

C.2 Pharmaceutical Case Study

Table C.2 shows results of the Durbin-Watson tests for the model fitted in section 4.4.4.2. As in Table C.1, there is insufficient evidence to reject the null hypothesis of no autocorrelation in the residuals. Therefore, the fitted model meets the assumptions of uncorrelated residuals required for fitting the VAR model. Similar results are observed for the other models fitted during the pharmaceutical case study.

Table C.2: Results of Durbin-Watson test for autocorrelation in the residuals. Panel A denotes the models fitted during the recessionary period (01/12/2007 to 30/06/2009), and Panel B denotes the models fitted during the non-recessionary (expansionary) period (01/07/2009 to 31/12/2015). The negative column denotes the VAR model fitted using the frequency of negative terms as a proxy of investor sentiment. The vice column denotes the equivalent model using the frequency of vice terms.

Panel A: Recession Period		
	Negative	Vice
Durbin-Watson	2.00	2.00
p -value	0.98	0.95
Panel B: Non-Recessionary Period		
	Negative	Vice
Durbin-Watson	2.00	2.00
p -value	0.92	0.99

Appendix D

GI Negative and Vice Terms

This section lists the negative and vice terms in the GI dictionary, together with their ranked frequencies in each of the corpora in this thesis. Ambiguous terms are marked with a # symbol followed by a unique number (e.g. #1, #2, #3) to denote the sense number of a particular word type. The GI's disambiguation rules (Kelly & Stone (1975)) are designed to analyse the context around an ambiguous term, and assign the most likely sense. Different sense numbers result in the assignment of different GI tags, and hence while certain senses of a given word might be tagged as negative or vice, others might not. This is important to note as it may not be immediately apparent why a particular word in this list should be tagged as negative or vice. An example of this is the entry 'Get#5', which is tagged as negative in the GI. A look at the GI's disambiguation rules reveals that this sense is only applied when the word appears as part of the idiomatic expression 'get rid of'. The disambiguation rules themselves number in the thousands, and will not be printed here due to space constraints. However, the interested reader should consider the work by Kelly & Stone (1975), which, in addition to describing their construction, lists the rules in their entirety.

Table D.1: GI negative and vice terms with frequency ranks in each corpus. The first column denotes the sense entry in the GI dictionary (with #1, #2, #3 etc. denoting the sense number if the specified sense is ambiguous.) The second column denotes the associated category with each sense, either negative, vice or both (negative and vice). The third column denotes the overall rank of the sense summed over all three corpora. The fourth column denotes the ranked frequency in the market level study using the Abreast of the Market corpus. The fifth column denotes the ranked frequency of the sense in the firm level study using the Enron corpus. The final column denotes the ranked frequency of the sense in industry level study using the pharmaceutical corpus.

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Cost#1	Negative	1	31	4	2
Loss	Both	2	4	2	5
Disease	Both	3	162	120	1
Expense#1	Negative	4	108	21	3
Cancer	Both	5	163	114	4
Lost#1	Negative	6	1	13	49
Lower#1	Negative	7	5	19	11
Drop#4	Negative	8	2	39	126
Close#2	Negative	9	6	12	19
Decline#2	Negative	10	3	14	33
Decline#1	Negative	11	7	32	28
Exclude	Negative	12	94	46	6
Concern#2	Negative	13	8	9	31
Against	Negative	14	18	7	9
Adverse	Both	15	180	53	7
Low#1	Negative	16	9	18	26
Need#1	Negative	17	43	10	10
Liability	Negative	18	137	30	8
Need#2	Negative	19	125	18	12
Cut#2	Negative	20	11	15	90
Negative	Both	21	30	51	22
Pain	Negative	22	170	106	13
Decrease#1	Negative	23	179	98	14
Disorder	Both	24	217	122	15
Chronic	Both	25	216	123	16
Inflation	Negative	26	10	98	175
Drive#2	Negative	27	87	100	18
Severe	Both	28	84	78	19
Infection	Negative	29	213	123	17
Charge#6	Negative	30	73	54	21
Problem	Negative	31	26	6	40
Charge#2	Negative	32	80	33	23
Rumor	Negative	33	12	81	213
Symptom	Both	34	212	122	20
Competition	Negative	35	79	20	27
Challenge#1	Negative	36	157	70	24
Block#1	Negative	37	13	68	112
Hit#1	Negative	38	14	44	117
Decrease#2	Negative	39	187	107	25
Weakness	Vice	40	15	105	89
Suffer#1	Negative	41	37	48	41
Division	Negative	42	72	27	35
Slight#3	Negative	43	19	84	98
Limit#3	Negative	44	78	5	42
Differ	Negative	45	198	20	30

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Competitive	Negative	46	111	17	36
Bankruptcy	Both	47	112	1	240
Too#2	Vice	48	25	24	130
Failure	Both	49	127	23	36
Worry#1	Negative	50	16	67	250
Limit#2	Negative	51	101	38	37
Drop#3	Negative	52	21	65	150
Fear#1	Negative	53	20	52	196
Error	Both	54	197	88	29
Lose#1	Negative	55	29	37	103
Loser	Negative	56	17	108	271
Poor#1	Both	57	24	101	127
Death	Negative	58	142	108	34
Breakdown	Negative	59	202	117	32
Fail#2	Negative	60	41	29	63
Push#1	Negative	61	22	68	186
Recession	Negative	62	23	89	196
Depreciation	Negative	63	191	77	38
Deficit	Both	64	34	120	103
Unfavorable	Both	65	179	119	39
Aggressive	Negative	66	48	31	72
Bad	Both	67	35	39	136
Difficult	Both	67	64	35	57
Hurt#1	Negative	68	27	69	198
Competitor	Negative	69	110	45	46
Damage#1	Both	70	75	58	56
Slump	Negative	71	28	113	211
Concern#1	Negative	72	40	44	116
Complex	Negative	73	201	27	48
Avoid	Negative	74	83	67	54
Plaintiff	Negative	75	209	40	45
Mistake#1	Both	76	188	86	44
Cost#2	Negative	77	135	53	52
Hedge	Negative	78	97	11	102
Shot#1	Negative	79	32	111	250
Allege	Negative	80	118	28	64
Nervous	Negative	81	71	106	69
Disclaim	Negative	82	228	116	43
Lack#1	Both	83	70	70	80
Disappoint#2	Both	84	36	110	187
Sank	Negative	85	33	121	277
Eliminate	Negative	86	129	61	59
Delay#1	Negative	87	101	76	66
Headache	Both	88	219	117	47
Mean#2	Both	89	123	120	53
Cut#1	Negative	90	47	80	135
Execute	Negative	91	154	72	58
Rat	Both	92	227	123	50
Crude	Negative	93	55	16	239
Rival#1	Negative	94	86	8	192
Regardless	Negative	95	155	115	55
Collapse#1	Negative	96	128	3	275
Hard#1	Both	97	57	52	138
Fracture	Both	98	223	123	51
Crisis	Both	99	88	19	123

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Volatility	Negative	100	46	41	221
Short#5	Both	101	38	103	217
Excess	Negative	102	111	69	77
Low#2	Negative	103	45	84	180
Reject	Negative	104	67	87	124
Dispute#1	Negative	105	119	35	97
Deny	Negative	106	59	59	156
Disappointment	Negative	107	42	112	234
Depression#2	Both	108	200	123	61
Fight#2	Negative	108	120	93	83
Antitrust	Negative	109	113	86	88
Duty	Negative	110	211	70	68
Limitation	Both	110	222	108	60
Retreat#2	Negative	111	39	119	275
Weaken	Negative	112	56	104	159
Inadequate	Both	113	180	110	67
Inhibit	Negative	114	223	122	62
Compete	Negative	115	103	63	104
Relapse	Negative	116	225	123	62
Difficulty	Both	117	123	51	99
Break#1	Negative	118	60	75	169
Cheap	Both	119	52	77	238
Violation	Both	119	187	68	76
Breach	Negative	120	207	59	75
Crash#1	Vice	120	44	110	274
Miss#1	Negative	121	65	108	140
Infect	Negative	122	218	122	65
Resistance	Vice	123	155	118	73
Commit#1	Negative	124	163	92	82
Drop#1	Negative	125	49	101	240
Contend	Negative	126	50	101	242
Threat	Both	127	112	94	102
War	Negative	128	53	98	240
Intervention	Both	129	179	116	74
Worst	Both	130	57	88	212
Abnormal	Both	131	215	119	70
Sluggish	Both	132	51	123	234
Force#3	Negative	133	81	26	201
Trouble#1	Both	134	68	47	209
Interfere	Negative	135	219	119	71
Delay#2	Negative	136	106	99	113
Close#7	Negative	137	85	76	143
Injury	Both	138	215	122	73
Withdrawn	Vice	138	200	103	83
Complaint	Negative	139	202	55	95
Beat#2	Negative	140	71	108	166
Volatile	Negative	140	62	87	222
Push#2	Negative	141	61	86	227
Burden#1	Both	142	199	108	84
Rash	Both	143	211	121	78
Unable	Vice	144	136	72	109
Retreat#1	Negative	145	54	115	279
Defensive	Negative	146	58	111	239
Complication	Both	147	223	119	81
Fail#1	Negative	147	130	43	132

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Fatigue	Negative	148	226	123	79
Retire	Negative	149	151	72	110
Omission	Negative	150	208	121	85
Fever	Both	151	177	123	93
Vomit	Negative	152	225	123	84
Emergency	Both	153	202	96	96
Attack#3	Both	154	207	118	91
Flu	Vice	155	219	122	87
Lag	Negative	156	66	116	224
Shortcoming	Both	156	219	122	88
Veto	Negative	157	220	122	88
Antagonist	Negative	158	227	123	86
Fraud	Negative	159	186	67	111
Barrier	Negative	160	145	115	107
Strain#1	Negative	161	210	121	94
Worry#2	Negative	162	63	105	270
Worry#4	Negative	163	69	83	259
Illness	Both	164	220	123	92
Disposal	Negative	165	195	113	101
Drag	Negative	165	74	85	252
Expensive	Both	166	96	88	171
Passe	Negative	167	109	93	149
Scheme#1	Negative	167	212	103	100
Vulnerable	Vice	167	89	115	168
Dominate#1	Negative	168	82	94	210
Worsen	Negative	169	158	104	118
Need#3	Negative	170	160	87	125
Insufficiency	Negative	171	228	122	95
Conflict#1	Negative	172	186	34	144
Shake#1	Negative	172	93	92	186
Depress#1	Negative	173	76	99	244
Expose	Negative	174	180	85	119
Unemployment	Vice	175	77	105	232
Worse	Both	175	95	85	187
Default	Negative	176	160	36	156
Struggle#2	Negative	176	104	91	165
Severity	Both	177	213	123	102
Charge#3	Negative	178	167	107	121
Infringement	Both	179	202	123	105
Exit	Negative	180	179	97	122
Fight#1	Negative	180	139	97	135
Absence	Negative	181	134	101	136
Restrict#1	Negative	182	165	104	128
Omit	Negative	183	170	122	117
Suspend	Negative	184	110	100	163
Owe	Negative	185	194	25	170
Battle#1	Negative	186	105	76	191
False	Both	187	182	91	129
Defendant	Negative	188	214	62	131
Fatal	Negative	189	222	114	106
Orphan	Negative	190	226	117	106
Abuse#1	Both	191	220	113	112
Dependence	Vice	192	225	119	108
Suspension	Negative	193	217	117	115
Block#2	Negative	194	137	113	147

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Rigorous	Both	195	221	120	114
Capture#2	Negative	196	125	113	163
Anxiety	Negative	197	142	115	148
Attack#1	Both	197	176	93	138
Doubt#1	Negative	198	117	74	213
Blame#2	Negative	199	90	89	262
Compel#2	Negative	200	179	100	137
Disappoint#1	Negative	201	125	107	176
Inhibition	Both	202	228	123	120
Wrong#1	Vice	202	121	57	233
Threaten	Negative	203	136	89	175
Allegation	Negative	204	183	91	145
Devastate	Negative	205	202	102	134
Refuse#1	Negative	205	156	62	184
Fat#1	Negative	206	189	121	133
Hostile	Negative	207	91	118	260
Cold#1	Negative	208	174	109	149
Dismiss#2	Negative	209	166	94	160
Fear#2	Negative	209	100	96	250
Cancel	Negative	210	137	64	216
Quit	Negative	211	208	82	147
Restriction	Negative	211	164	115	152
Erosion	Negative	212	192	106	146
Erase	Negative	213	92	114	277
Junk	Negative	213	171	22	280
Divide#1	Negative	214	146	108	176
Argue#1	Negative	215	99	95	274
Illegal	Both	216	208	77	156
Elimination	Negative	217	208	108	142
Mislead	Negative	217	192	105	151
Shortage	Negative	218	183	95	162
Impair	Negative	219	223	105	139
Suspect#2	Negative	219	180	114	154
Challenge#2	Negative	220	182	90	169
Violate	Negative	220	173	88	177
Sag	Negative	221	98	117	275
Criminal	Negative	222	187	98	165
Dependent	Negative	222	193	102	158
Cool	Negative	223	114	115	237
Oversight	Both	223	225	66	167
Withhold	Negative	224	221	109	143
Trouble#3	Negative	225	133	57	267
Chase#2	Negative	226	102	109	273
Defect	Both	226	216	120	141
Injunction	Negative	227	199	99	162
Slash	Negative	228	113	107	253
Temporarily	Negative	228	151	96	201
Waste#1	Both	229	143	105	203
Addiction	Negative	230	227	123	138
Disruption	Both	231	210	111	153
Unlawful	Negative	231	227	118	141
Capture#1	Negative	232	178	112	172
Excessive	Both	233	169	117	178
Prohibit	Negative	234	208	76	181
Lack#2	Negative	235	161	114	189

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Marginal	Negative	236	141	122	202
Stress#1	Negative	236	205	115	157
Kill#1	Negative	237	158	99	204
Epidemic	Negative	238	228	123	145
Inability	Both	239	181	114	176
Interruption	Negative	240	200	118	164
Nervousness	Negative	241	115	117	265
Absent#1	Negative	242	187	103	185
Disappoint#3	Negative	243	126	110	250
Pessimism	Negative	244	107	121	282
Woe	Negative	244	140	71	267
Hunter	Negative	245	180	119	182
Wound#3	Both	245	220	119	155
Casualty	Negative	246	120	108	268
Dangerous	Both	246	159	121	200
Falter	Negative	246	114	112	282
Disaster	Both	247	162	104	213
Harm#1	Both	247	218	118	161
Painful	Both	247	193	120	175
Damage#2	Negative	248	174	91	214
Dump#2	Negative	249	122	99	279
Panic	Negative	249	116	118	275
Collapse#2	Negative	250	156	56	280
Impose	Negative	250	188	95	200
Wild#1	Negative	251	128	107	258
Strike#1	Negative	252	128	112	254
Danger	Both	253	152	111	223
Blindness	Both	254	228	123	155
Turmoil	Both	254	130	102	260
Confusion	Both	255	144	105	238
Secret	Negative	255	213	95	185
Erode	Negative	256	127	116	258
Improper	Both	256	207	109	181
Incredible	Negative	256	187	110	194
Dominant	Vice	257	166	79	242
Abandon	Negative	258	144	88	261
Compel#1	Negative	258	204	121	178
Badly	Both	259	131	98	271
Mar#1	Both	259	226	122	164
Questionable	Both	259	208	42	264
Deficiency	Both	260	206	122	177
Lost#3	Both	261	139	114	247
Loom	Negative	262	149	106	244
Reluctant	Negative	263	134	103	270
Exclusion	Negative	264	225	123	168
Undue	Both	264	210	122	179
Shock#1	Negative	265	141	103	260
Bar#4	Negative	266	183	103	218
Pessimistic	Negative	267	124	121	279
Complexity	Negative	268	225	101	188
Enforce	Negative	268	222	116	179
Poor#6	Both	269	156	111	240
Irritation	Negative	270	226	123	173
Overwhelming	Negative	270	177	114	219
Fear#3	Negative	271	226	123	174

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Restrict#2	Negative	271	220	121	181
Insufficient	Both	272	222	113	186
Liquidation	Both	272	184	94	232
Doubt#2	Negative	273	143	103	270
Incorrect	Both	273	188	112	215
Skeptical	Both	274	138	109	274
Rejection	Both	275	207	122	193
Costly	Both	276	203	101	215
Discrimination	Vice	276	226	123	180
Harm#2	Negative	276	217	120	189
Bound#3	Both	277	206	119	198
Charge#7	Negative	277	180	99	237
Complain	Negative	277	156	94	268
Defective	Both	277	223	121	184
Unfortunate#2	Both	277	208	111	204
Isolate#1	Negative	278	218	121	190
Artificial	Both	279	205	103	215
Bout	Negative	280	132	121	278
Substitution	Negative	280	226	123	183
Outbreak	Negative	281	199	122	207
Penalty	Vice	282	217	108	204
Unexpectedly	Both	283	146	109	274
Discourage#2	Negative	284	152	116	260
Dispense	Negative	284	227	123	186
Risky	Both	284	173	98	253
Wary	Negative	285	150	99	281
Accuse#2	Negative	286	207	73	250
Depress#3	Negative	286	139	121	275
Anomaly	Negative	287	214	123	197
Demise	Negative	287	214	50	279
Blow#1	Negative	288	175	94	259
Discomfort	Negative	288	221	121	195
Abrupt	Negative	289	165	109	257
Complicate#2	Negative	290	211	98	224
Fire#1	Negative	290	196	90	244
Laid#2	Negative	290	217	49	280
Oust	Negative	291	212	59	273
Sore	Negative	291	221	122	196
Scar	Negative	292	220	119	200
Undermine	Negative	292	188	98	246
Bother#1	Negative	293	182	123	229
Resignation	Negative	293	191	88	254
Trouble#2	Negative	293	157	108	270
Destroy	Negative	294	207	105	225
Gloomy	Negative	294	141	121	279
Explosive	Negative	295	148	113	279
Suppression	Negative	295	228	123	193
Contrary	Negative	296	182	116	238
Dismal	Both	296	160	119	259
Spill	Negative	296	152	115	273
Tension	Negative	296	161	121	256
Dispose#2	Negative	297	210	98	230
Guilty	Both	297	205	116	220
Hazard#1	Both	297	223	118	203
Kick#1	Negative	297	147	121	275

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Liquidate	Negative	297	172	87	278
Trauma	Both	297	223	122	199
Argument#1	Negative	298	182	107	249
Dull	Negative	298	147	120	277
Rough#1	Negative	298	156	113	271
Burn#2	Negative	299	201	86	252
Disrupt	Negative	299	196	114	230
Fret	Negative	299	140	123	283
Knock#3	Negative	299	142	123	281
Obstacle	Negative	299	183	114	242
Suspect#1	Negative	299	183	118	238
Bail	Negative	300	156	105	281
Suicide	Vice	300	224	120	203
Bankrupt	Both	301	223	60	271
Fallout	Negative	301	178	85	278
Fine#6	Negative	301	219	108	218
Struggle#1	Negative	301	173	114	254
Terrible	Both	301	172	114	255
Complicate#1	Negative	302	214	100	229
Invalid	Both	302	223	121	205
Punish	Negative	302	169	101	273
Shrug#2	Negative	302	148	117	282
Suppress	Negative	302	222	121	206
Desperate	Negative	303	203	106	234
Divert	Negative	303	217	123	209
Hunt#2	Negative	303	156	118	271
Stress#2	Negative	303	203	112	229
Boast	Both	304	217	110	221
Poison	Negative	304	201	122	224
Deadly	Both	305	222	120	209
Agitation	Both	306	228	123	202
Distress#1	Negative	306	206	104	236
Oppose#1	Negative	306	186	97	263
Dark	Negative	307	205	107	235
Instability	Both	307	208	115	227
Unclear	Both	307	212	89	246
Jobless	Negative	308	162	117	271
Beat#3	Both	309	156	119	276
Combat#1	Negative	309	212	121	220
Combat#2	Negative	309	227	123	206
Criticize	Negative	310	206	99	245
Malignant	Negative	310	226	123	208
Scandal	Negative	310	167	106	279
Bruise	Negative	311	203	122	228
Cancellation	Negative	312	204	110	238
Deviation	Negative	312	226	123	209
Hamper	Negative	312	190	114	248
Pound#2	Negative	312	178	119	255
Struck#1	Negative	312	180	101	271
Knock#1	Negative	313	153	120	282
Misuse	Both	313	225	123	211
Plague#2	Negative	313	169	115	270
Seize	Negative	313	198	107	248
Curtail	Negative	314	175	112	267
Dim#1	Negative	314	177	118	259

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Gloom	Negative	314	156	117	283
Vain	Both	314	224	122	213
Wane	Negative	314	167	116	273
Languish	Negative	315	157	122	278
Controversial	Negative	316	206	80	272
Jittery	Negative	316	168	118	271
Force#6	Both	317	205	96	256
Leak	Negative	318	203	118	237
Opposition	Negative	318	197	107	254
Tariff	Negative	318	203	77	280
Trap#1	Negative	318	211	123	227
Bleak	Both	319	168	121	271
Interference	Both	319	227	120	216
Lapse	Negative	319	219	120	223
Withheld	Negative	319	222	108	230
Accident	Negative	320	192	120	248
Overlook	Negative	320	188	114	258
Contamination	Both	321	224	123	218
Dead	Negative	321	170	114	277
Poor#2	Both	321	198	117	246
Prod	Negative	321	223	117	224
Susceptible	Negative	321	213	122	228
Exhaust	Negative	322	195	116	251
Exploit#2	Negative	322	196	116	250
Unfair	Both	322	215	106	241
Hunger	Negative	323	225	123	219
Interrupt	Negative	323	204	120	239
Constrain	Negative	324	207	121	236
Lure	Negative	324	169	117	279
Refusal	Negative	324	211	111	242
Unsuccessful	Both	324	190	119	255
Unsure	Vice	324	221	110	233
Constraint	Negative	325	216	121	229
Critic	Negative	325	206	87	272
Tax#2	Negative	325	200	122	243
Thwart	Negative	325	174	120	271
Disproportionate	Negative	326	213	123	231
Downfall	Negative	326	226	67	280
Lay#3	Negative	326	181	110	275
Sick#1	Both	326	208	122	236
Harmful	Both	327	218	118	232
Controversy	Negative	328	207	90	271
Oppose#2	Negative	328	206	117	245
Poor#4	Both	328	224	120	227
Poverty	Both	328	228	120	223
Break#5	Negative	329	206	122	241
Flaw	Negative	329	208	111	250
Irregular	Negative	329	225	123	224
Sour	Both	329	178	117	274
Blind#1	Negative	330	222	118	231
Burn#1	Both	330	216	120	234
Grab	Negative	330	194	109	267
Rattle#2	Negative	330	172	115	283
Discharge#1	Negative	331	228	123	223
Gamble#2	Negative	331	176	115	280

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Novice	Negative	331	228	123	223
Overturn	Negative	331	208	119	244
Attack#2	Negative	332	216	115	241
Blast#1	Negative	332	209	119	244
Blame#1	Both	333	203	98	272
Confuse#3	Both	333	177	120	276
Injure	Negative	333	222	120	232
Neglect#1	Negative	333	227	123	226
Strict#1	Both	333	219	110	244
Turbulent	Negative	333	188	120	265
Battle#2	Negative	334	207	120	247
Confront	Negative	334	207	120	247
Explode	Negative	334	178	119	277
Temper#2	Negative	334	179	123	272
Isolate#2	Negative	335	211	122	242
Neutralize	Negative	335	223	123	230
Scare#1	Negative	335	182	121	272
Struggle#3	Negative	335	199	113	263
Unnecessary	Both	335	223	118	234
Arrest#1	Negative	336	223	119	234
Deceptive	Negative	336	224	123	230
Protest#1	Negative	336	223	113	240
Stringent	Negative	336	223	115	238
Succumb	Negative	336	178	121	277
Ail	Negative	337	193	111	273
Confuse#4	Both	337	186	116	275
Empty#1	Negative	337	216	116	245
Hazardous	Both	337	204	121	252
Hostility	Negative	337	205	120	252
Relentless	Vice	337	201	120	256
Revolution	Negative	337	220	108	249
Scared#1	Negative	337	184	115	278
Suffer#3	Negative	337	216	120	241
Unnerve	Negative	337	179	116	282
Unwilling	Negative	337	200	115	262
Expense#2	Negative	338	202	117	259
Explosion	Negative	338	187	117	274
Objection	Negative	338	217	111	250
Anxious#2	Negative	339	190	118	271
Congestion	Negative	339	228	117	234
Faint	Negative	339	223	122	234
Radical	Both	339	208	112	259
Afflict	Negative	340	216	123	241
Hinder	Negative	340	211	121	248
Insignificant	Both	340	223	119	238
Manipulation	Negative	340	222	91	267
Disadvantage	Both	341	224	122	235
Naive	Both	341	223	122	236
Sap	Negative	341	187	122	272
Vague	Negative	341	184	118	279
Fraudulent	Negative	342	220	120	242
Ill#1	Both	342	214	123	245
Revoke	Negative	342	223	122	237
Unforeseen	Negative	342	220	118	244
Disagreement	Negative	343	211	117	255

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Discourage#1	Negative	343	196	122	265
Standstill	Negative	343	195	123	265
Subside	Negative	343	185	120	278
Crumble	Negative	344	197	105	282
Doldrums	Both	344	180	122	282
Empty#2	Negative	344	228	117	239
Irregularity	Negative	344	206	101	277
Stigma	Both	344	228	122	234
Blind#2	Negative	345	228	123	234
Counterfeit	Both	345	227	123	235
Depression#1	Negative	345	228	123	234
Destruction	Both	345	224	115	246
Shun	Negative	345	184	121	280
Tired#1	Negative	345	195	123	267
Unpredictable	Vice	345	213	121	251
Bite#1	Negative	346	211	120	255
Disable	Negative	346	227	123	236
Distraction	Both	346	226	120	240
Doubtful	Both	346	218	115	253
Grim	Both	346	186	119	281
Unstable	Both	346	217	122	247
Abate	Negative	347	195	122	270
Cripple	Negative	347	206	102	279
Disturbance	Both	347	226	123	238
Dominate#2	Negative	347	223	123	241
Prejudice	Negative	347	227	120	240
Refrain	Negative	347	211	110	266
Wild#4	Negative	347	192	115	280
Avert	Negative	348	207	113	268
Broke#3	Both	348	203	120	265
Cramp	Negative	348	228	123	237
Crush#2	Negative	348	203	119	266
Distort	Negative	348	197	121	270
Dwindle	Negative	348	200	116	272
Hectic	Both	348	186	123	279
Implicate	Negative	348	226	122	240
Ironic	Both	348	203	118	267
Misrepresent	Negative	348	227	111	250
Rage	Negative	348	210	121	257
Blur	Negative	349	226	123	240
Deter	Negative	349	206	117	266
Dispute#2	Negative	349	207	103	279
Fragile	Negative	349	201	116	272
Rigor	Negative	349	227	123	239
Boot	Negative	350	221	123	246
Exempt	Negative	350	220	117	253
Jeopardy	Negative	350	198	114	278
Make#2	Negative	350	210	117	263
Manipulate	Negative	350	218	111	261
Abuse#2	Negative	351	225	121	245
Confine#2	Negative	351	192	121	278
Desire#3	Negative	351	224	123	244
Pollution	Both	351	197	116	278
Troublesome	Both	351	200	123	268
Use#6	Vice	351	221	122	248

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Dictate	Negative	352	203	121	268
Dismiss#1	Negative	352	209	119	264
Furious	Negative	352	195	114	283
Indefinite	Both	352	199	122	271
Lose#2	Negative	352	207	112	273
Prosecution	Negative	352	228	123	241
Shoot#1	Negative	352	192	118	282
Sneak	Negative	352	205	123	264
Startle	Negative	352	198	119	275
Conflict#2	Negative	353	194	120	279
Gun	Negative	353	198	117	278
Hesitant	Vice	353	194	121	278
Killer	Negative	353	226	123	244
Meager	Negative	353	197	122	274
Punch	Negative	353	197	114	282
Stray	Negative	353	191	123	279
Depress#2	Both	354	194	121	279
Divide#2	Negative	354	220	121	253
Flee	Negative	354	202	114	278
Fought	Negative	354	204	117	273
Scary	Negative	354	194	122	278
Void	Both	354	223	102	269
Wrinkle	Both	354	221	123	250
Aggravate	Negative	355	218	122	255
Clog	Negative	355	225	123	247
Denial	Negative	355	210	115	270
Disturb#2	Both	355	201	123	271
Dump#1	Negative	355	194	119	282
Hurt#2	Negative	355	192	120	283
Liable	Negative	355	218	121	256
Restless	Negative	355	224	123	248
Sad	Negative	355	224	111	260
Violence	Both	355	222	121	252
Conceal	Negative	356	225	111	260
Confound	Negative	356	204	123	269
Fighter	Negative	356	205	121	270
Fled	Negative	356	202	114	280
Irritable	Negative	356	228	123	245
Paralysis	Negative	356	220	123	253
Scare#2	Negative	356	202	117	277
Strict#2	Both	356	218	119	259
Uncomfortable	Negative	356	202	118	276
Uneasy	Negative	356	192	122	282
Contempt	Both	357	228	87	282
Defeat#1	Both	357	208	118	271
Deplete	Negative	357	213	119	265
Disregard	Negative	357	220	123	254
Diversion	Negative	357	223	117	257
Gamble#1	Negative	357	203	117	277
Neglect#2	Negative	357	217	123	257
Scrutinize	Negative	357	206	113	278
Ugly	Both	357	198	122	277
Upset#1	Negative	357	216	115	266
Violent	Both	357	194	121	282
Atrophy	Negative	358	227	123	248

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Fire#2	Negative	358	227	101	270
Formidable	Both	358	214	114	270
Frustration	Negative	358	212	109	277
Nasty	Negative	358	193	122	283
Ominous	Both	358	198	120	280
Rebuff	Negative	358	201	118	279
Recede	Negative	358	196	121	281
Revert	Negative	358	219	123	256
Unattractive	Both	358	197	123	278
Unfortunate#1	Both	358	224	106	268
Defy	Negative	359	198	120	281
Harsh	Both	359	207	121	271
Haunt	Negative	359	202	119	278
Parasite	Negative	359	227	123	249
Stolen#1	Both	359	225	121	253
Tragedy	Negative	359	220	102	277
Alarming	Both	360	215	121	264
Bitter	Negative	360	222	98	280
Fierce	Negative	360	211	121	268
Fleeting	Negative	360	195	122	283
Noise	Negative	360	205	119	276
Rampant	Both	360	205	117	278
Sickness	Both	360	226	123	251
Superficial	Both	360	227	123	250
Vice#1	Both	360	227	119	254
Avoidance	Negative	361	227	121	253
Defeat#2	Negative	361	209	122	270
Dire	Both	361	207	118	276
Fire#3	Negative	361	222	105	274
Frivolous	Negative	361	227	101	273
Ridiculous	Both	361	206	113	282
Sputter	Negative	361	199	119	283
Waste#2	Negative	361	217	122	262
Accusation	Negative	362	222	104	276
Blunt	Negative	362	205	120	277
Catastrophe	Both	362	218	118	266
Disastrous	Both	362	205	119	278
Eradicate	Negative	362	226	123	253
Hesitation	Vice	362	198	122	282
Lack#3	Both	362	216	121	265
Stain	Negative	362	227	123	252
Sufferer	Negative	362	228	123	251
Unsettling	Negative	362	197	123	282
Ax	Negative	363	228	121	254
Bloody	Negative	363	211	122	270
Crazy	Both	363	201	120	282
Deprive	Negative	363	221	122	260
Involuntary	Negative	363	225	122	256
Irrational	Both	363	207	114	282
Precipitate	Negative	363	210	121	272
Slam	Negative	363	206	118	279
Tired#2	Negative	363	218	122	263
Traumatic	Negative	363	224	123	256
Trouble#4	Negative	363	209	119	275
Diseased	Both	364	226	123	255

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Displace	Negative	364	225	118	261
Dissent	Both	364	225	123	256
Dissolution	Negative	364	225	120	259
Drowsiness	Negative	364	228	123	253
Inflict	Negative	364	210	121	273
Mess#1	Both	364	218	109	277
Nightmare	Negative	364	220	110	274
Raid	Negative	364	207	121	276
Stifle	Negative	364	207	119	278
Unpleasant	Both	364	207	118	279
Endanger	Negative	365	220	112	273
Fault	Both	365	223	117	265
Fearful	Negative	365	207	116	282
Frustrate#3	Negative	365	206	121	278
Obstruction	Negative	365	228	123	254
Regression	Both	365	228	122	255
Slight#2	Negative	365	204	121	280
Contagious	Negative	366	226	123	257
Dispel	Negative	366	209	122	275
Disturb#3	Negative	366	209	121	276
Dizzy	Negative	366	213	118	275
Frustrate#2	Negative	366	212	115	279
Ineffective	Negative	366	225	122	259
Jeopardize	Negative	366	216	117	273
Obscure	Negative	366	218	112	276
Unemployed	Both	366	227	118	261
Unprofitable	Negative	366	210	121	275
Fall#7	Negative	367	218	107	282
Hardship	Both	367	227	123	257
Limp	Negative	367	205	119	283
Precarious	Both	367	219	105	283
Regret#1	Negative	367	225	112	270
Shame	Negative	367	226	115	266
Static	Both	367	222	118	267
Terrorism	Negative	367	223	121	263
Weed	Negative	367	215	122	270
Alarm#2	Negative	368	208	118	282
Apprehension	Negative	368	214	122	272
Bore#6	Both	368	207	121	280
Choke#2	Negative	368	206	123	279
Deficient	Both	368	228	122	258
Dent	Negative	368	220	114	274
Detrimental	Both	368	219	122	267
Dim#2	Negative	368	205	123	280
Din	Negative	368	222	121	265
Futility	Both	368	227	123	258
Infiltration	Both	368	228	123	257
Monster	Negative	368	220	123	265
Muddle	Negative	368	203	122	283
Object#3	Negative	368	221	120	267
Stark	Negative	368	220	116	272
Theft	Negative	368	227	117	264
Ache	Negative	369	228	123	258
Capitulate	Negative	369	206	120	283
Contaminate	Negative	369	221	120	268

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Crushing	Negative	369	217	120	272
Discriminate	Negative	369	220	123	266
Disinterest	Both	369	224	123	262
Frighten#2	Negative	369	206	122	281
Idle#1	Vice	369	211	117	281
Impede	Negative	369	224	122	263
Lie#3	Negative	369	211	122	276
Lull	Negative	369	206	123	280
Paltry	Negative	369	206	123	280
Pinch#1	Both	369	216	123	270
Reckless	Both	369	225	121	263
Shallow	Negative	369	208	120	281
Shock#2	Negative	369	213	116	280
Sin#1	Both	369	220	121	268
Stagnant	Negative	369	216	123	270
Tamper	Negative	369	225	123	261
Tolerable	Negative	369	227	122	260
Topple	Negative	369	204	123	282
Vanish	Negative	369	216	110	283
Vulnerability	Vice	369	211	123	275
Aggression	Both	370	227	118	265
Bomb#1	Negative	370	210	118	282
Confuse#1	Negative	370	215	121	274
Crave	Negative	370	225	122	263
Crush#1	Negative	370	223	123	264
Daunting	Negative	370	224	121	265
Dilemma	Both	370	211	121	278
Disingenuous	Negative	370	228	103	279
Dispose#1	Negative	370	226	123	261
Hell	Negative	370	210	123	277
Murky	Negative	370	222	110	278
Negligence	Both	370	227	123	260
Notorious	Both	370	216	116	278
Plight	Negative	370	224	110	276
Restrict#3	Both	370	216	120	274
Worthless	Both	370	220	108	282
Assault#1	Both	371	209	121	281
Chaos	Both	371	215	118	278
Debtor	Negative	371	222	108	281
Frighten#1	Negative	371	209	123	279
Get#5	Negative	371	213	121	277
Lurk	Negative	371	211	120	280
Odd	Both	371	213	121	277
Provoke	Negative	371	218	112	281
Unwillingness	Negative	371	219	119	273
Whip#3	Negative	371	209	122	280
Wrongful	Both	371	228	108	275
Belie	Negative	372	208	123	281
Counteract	Negative	372	221	121	270
Devil	Negative	372	224	123	265
Discharge#2	Negative	372	227	120	265
Frighten#3	Negative	372	210	122	280
Frustrate#1	Negative	372	221	116	275
Ill#2	Both	372	222	118	272
Impediment	Both	372	224	120	268

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Nuts	Negative	372	215	123	274
Opponent	Negative	372	222	112	278
Poor#5	Negative	372	223	120	269
Reactive	Negative	372	227	123	262
Unhappy	Negative	372	216	115	281
Upheaval	Negative	372	213	122	277
Wrong#3	Both	372	226	110	276
Bolt	Negative	373	210	123	280
Clash	Negative	373	216	116	281
Droop	Negative	373	223	123	267
Err	Negative	373	221	116	276
Fake	Negative	373	224	122	267
Frantic	Negative	373	210	120	283
Impasse	Negative	373	210	122	281
Inefficiency	Negative	373	225	118	270
Knock#2	Negative	373	211	120	282
Lie#2	Negative	373	219	119	275
Sting	Negative	373	225	120	268
Stuffy	Both	373	227	123	263
Tear#1	Negative	373	218	121	274
Trample	Negative	373	218	118	277
Undesirable	Both	373	228	123	262
Unhealthy	Negative	373	222	123	268
Unreasonable	Negative	373	218	120	275
Wrestle	Negative	373	210	122	281
Ailment	Negative	374	224	123	267
Backward	Negative	374	222	121	271
Blind#4	Both	374	228	121	265
Contradict	Negative	374	220	121	273
Darken	Negative	374	215	121	278
Dissatisfaction	Both	374	223	122	269
Feeble	Both	374	212	122	280
Grapple	Negative	374	214	119	281
Hobble	Negative	374	209	122	283
Irresponsible	Both	374	224	123	267
Mediocre	Negative	374	219	123	272
Mistaken	Negative	374	219	123	272
Poor#3	Both	374	221	123	270
Runaway	Negative	374	216	119	279
Sloppy	Negative	374	209	122	283
Smash#1	Negative	374	212	121	281
Sorry#1	Negative	374	221	122	271
Subtract	Negative	374	218	119	277
Apprehensive	Negative	375	217	123	275
Captive	Negative	375	224	120	271
Corrupt	Both	375	225	119	271
Culprit	Both	375	218	122	275
Distress#2	Negative	375	223	116	276
Drought	Negative	375	212	123	280
Embarrassment	Both	375	226	114	275
Erroneous	Negative	375	221	121	273
Grave#1	Negative	375	227	123	265
Hoard	Negative	375	212	122	281
Hungry	Negative	375	221	121	273
Invade	Negative	375	210	123	282

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Leakage	Negative	375	225	122	268
Needy	Negative	375	228	121	266
Pest	Both	375	228	119	268
Shrug#1	Negative	375	211	123	281
Suspicious	Negative	375	220	122	273
Trick#1	Both	375	210	123	282
Unrest	Both	375	215	123	277
Vicious	Both	375	215	121	279
Angry	Negative	376	222	114	280
Beset	Negative	376	213	121	282
Broke#7	Negative	376	218	120	278
Crime	Both	376	225	120	271
Destructive	Both	376	227	123	266
Disturb#1	Negative	376	217	120	279
Flounder	Negative	376	213	121	282
Frown#1	Negative	376	228	123	265
Glum	Negative	376	211	122	283
Hate#1	Negative	376	212	123	281
Inconsistency	Negative	376	227	122	267
Inferior	Both	376	228	123	265
Redundant	Both	376	228	116	272
Repeal	Negative	376	222	118	276
Rival#2	Negative	376	214	122	280
Sickly	Both	376	210	123	283
Strain#2	Negative	376	219	122	275
Suspicion	Negative	376	215	121	280
Temper#1	Negative	376	216	120	280
Treacherous	Both	376	210	123	283
Balk	Negative	377	214	120	283
Bewilder	Negative	377	222	112	283
Blatant	Negative	377	227	122	268
Bribe	Both	377	219	118	280
Bullet	Negative	377	218	121	278
Burden#2	Negative	377	218	122	277
Deluge	Negative	377	214	122	281
Dissatisfied	Negative	377	223	122	272
Distract	Negative	377	217	122	278
Fat#2	Negative	377	211	123	283
Horrible	Both	377	222	116	279
Indictment	Negative	377	216	122	279
Involve#5	Both	377	226	121	270
Misunderstood	Negative	377	228	116	273
Negligent	Both	377	226	121	270
Obsolete	Negative	377	224	119	274
Perplex	Negative	377	213	122	282
Refugee	Negative	377	226	122	269
Savage	Negative	377	222	113	282
Shove	Negative	377	211	123	283
Stalemate	Negative	377	214	122	281
Trap#2	Negative	377	213	123	281
Vengeance	Negative	377	212	122	283
Whack	Negative	377	211	123	283
Abandonment	Negative	378	227	122	269
Awful#1	Both	378	217	123	278
Blast#2	Negative	378	218	119	281

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Bury	Negative	378	217	120	281
Evade	Negative	378	225	123	270
Fabrication	Both	378	220	121	277
Indecision	Vice	378	212	123	283
Prohibitive	Both	378	226	123	269
Protest#2	Negative	378	225	120	273
Restlessness	Negative	378	226	123	269
Rupture#1	Both	378	226	123	269
Scream#2	Negative	378	216	121	281
Stricken	Negative	378	228	108	282
Taint	Negative	378	222	117	279
Thirst	Negative	378	223	123	272
Upset#5	Negative	378	216	119	283
Worry#3	Negative	378	213	123	282
Agonize	Negative	379	218	122	279
Arrest#2	Negative	379	222	118	279
Bothersome	Both	379	226	123	270
Censor	Negative	379	228	123	268
Creep#1	Vice	379	213	123	283
Deal#6	Negative	379	219	121	279
Deceive#2	Negative	379	226	123	270
Degrade	Negative	379	228	123	268
Delinquency	Both	379	222	123	274
Desperation	Negative	379	224	114	281
Doom#1	Both	379	215	122	282
Fool#2	Negative	379	217	121	281
Grave#2	Both	379	227	121	271
Miserable	Negative	379	217	122	280
Outsider	Negative	379	223	113	283
Overrun	Negative	379	221	121	277
Plague#1	Both	379	218	122	279
Ravage	Negative	379	223	123	273
Scarcity	Negative	379	218	121	280
Shock#4	Negative	379	222	119	278
Slap	Negative	379	219	121	279
Spoil	Negative	379	214	123	282
Stole	Negative	379	216	123	280
Strict#3	Both	379	228	115	276
Strife	Both	379	222	114	283
Stubbornly	Negative	379	214	122	283
Unbelievable	Both	379	219	120	280
Uneven	Negative	379	216	123	280
Unfriendly	Both	379	213	123	283
Unrealistic	Vice	379	217	123	279
Anger#1	Negative	380	226	116	278
Burdensome	Both	380	227	121	272
Detachment	Both	380	226	123	271
Deviate	Negative	380	223	123	274
Engulf	Negative	380	221	116	283
Fickle	Negative	380	215	122	283
Fool#1	Negative	380	219	121	280
Hate#2	Negative	380	219	118	283
Jail#1	Negative	380	227	114	279
Jar#2	Negative	380	216	121	283
Outrageous	Both	380	220	122	278

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Peril	Negative	380	219	120	281
Steal#1	Negative	380	224	118	278
Undo	Negative	380	215	122	283
Uneasiness	Negative	380	214	123	283
Unsafe	Negative	380	228	123	269
Wear#2	Negative	380	222	123	275
Yawn	Negative	380	220	123	277
Accuse#1	Negative	381	227	121	273
Beware	Negative	381	216	123	282
Broke#6	Both	381	225	118	278
Buckle	Negative	381	217	122	282
Craze	Both	381	215	123	283
Darkness	Negative	381	227	123	271
Distortion	Negative	381	221	120	280
Dying#1	Negative	381	224	123	274
Extinguish	Negative	381	220	123	278
Havoc	Both	381	218	122	281
Impatient	Both	381	218	120	283
Nag#2	Negative	381	217	123	281
Rigid	Both	381	223	120	278
Sucker	Negative	381	215	123	283
Torrent	Negative	381	218	122	281
Turn#8	Negative	381	218	121	282
Unavoidable	Negative	381	228	121	272
Unjust	Both	381	227	123	271
Unjustified	Both	381	225	122	274
Untrue	Both	381	220	120	281
Woeful	Negative	381	220	122	279
Wound#1	Negative	381	224	120	277
Anxious#3	Negative	382	219	123	280
Apathy	Both	382	219	123	280
Battlefield	Negative	382	227	122	273
Burn#3	Negative	382	227	118	277
Carry#5	Vice	382	217	122	283
Cynical	Both	382	223	123	276
Dearth	Both	382	222	122	278
Depreciate	Negative	382	216	123	283
Feverish	Negative	382	218	121	283
Fight#3	Both	382	223	121	278
Fine#7	Negative	382	222	118	282
Hollow	Negative	382	224	123	275
Hunt#1	Negative	382	219	121	282
Ill#3	Both	382	218	123	281
Incurable	Negative	382	227	123	272
Indecisive	Vice	382	216	123	283
Indifference	Negative	382	221	120	281
Indifferent	Negative	382	216	123	283
Inflame	Negative	382	227	122	273
Mar#2	Negative	382	227	123	272
Meaningless	Both	382	222	121	279
Mine#3	Negative	382	216	123	283
Oppression	Negative	382	228	123	271
Paralyzed	Negative	382	221	123	278
Prison#1	Negative	382	223	122	277
Ration	Negative	382	228	121	273

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Rid	Negative	382	224	123	275
Rude	Both	382	219	122	281
Ruin#2	Negative	382	223	118	281
Sort#3	Both	382	217	123	282
Stupid#1	Both	382	217	122	283
Aggrieve	Negative	383	228	119	276
Aimless	Negative	383	217	123	283
Bite#2	Negative	383	219	123	281
Blow#3	Negative	383	218	122	283
Chaotic	Both	383	222	118	283
Colony	Negative	383	227	123	273
Conspire	Negative	383	220	123	280
Dirty	Both	383	223	122	278
Disdain	Both	383	223	120	280
Doom#2	Negative	383	221	119	283
Dying#2	Negative	383	228	120	275
Embarrass	Negative	383	222	119	282
Hack	Both	383	227	120	276
Handicap	Both	383	223	120	280
Irony	Both	383	225	118	280
Lonely	Negative	383	221	120	282
Lost#2	Negative	383	220	122	281
Misunderstanding	Negative	383	227	117	279
Numb	Negative	383	225	123	275
Offensive	Both	383	219	121	283
Outrage	Negative	383	226	118	279
Perverse	Both	383	218	122	283
Plod	Negative	383	217	123	283
Quandary	Negative	383	217	123	283
Ragged	Vice	383	218	122	283
Rip#2	Negative	383	221	120	282
Shell#3	Negative	383	224	123	276
Shiver	Negative	383	223	122	278
Sketchy	Both	383	224	116	283
Smack	Negative	383	219	121	283
Wrong#2	Negative	383	218	122	283
Careen	Negative	384	220	121	283
Clamor	Both	384	219	122	283
Confrontation	Both	384	220	121	283
Confuse#2	Negative	384	224	123	277
Covet	Negative	384	225	122	277
Cumbersome	Negative	384	228	120	276
Decay#1	Negative	384	228	115	281
Delusion	Both	384	227	123	274
Deterrent	Negative	384	226	121	277
Dubious	Both	384	223	122	279
Glare	Negative	384	223	123	278
Harassment	Negative	384	226	119	279
Inaccessible	Both	384	228	123	273
Litter	Both	384	224	122	278
Misery	Negative	384	220	123	281
Peculiar	Both	384	226	122	276
Point#6	Negative	384	228	113	283
Puny	Both	384	218	123	283
Retaliate	Negative	384	224	122	278

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Scared#2	Negative	384	219	122	283
Smear	Negative	384	226	123	275
Soreness	Negative	384	228	123	273
Tense	Negative	384	224	121	279
Upset#2	Negative	384	218	123	283
Wallow	Negative	384	218	123	283
Agitate	Negative	385	227	121	277
Anger#2	Negative	385	226	119	280
Antagonism	Negative	385	228	123	274
Arbitrary	Both	385	225	123	277
Arrogance	Both	385	228	114	283
Beg	Negative	385	224	122	279
Commonplace	Negative	385	222	122	281
Corruption	Both	385	227	117	281
Denounce	Negative	385	228	122	275
Devastation	Both	385	225	122	278
Disbelief	Negative	385	219	123	283
Dishearten	Negative	385	221	121	283
Dreadful	Both	385	225	121	279
Enemy	Negative	385	226	118	281
Envy	Both	385	227	118	280
Equivocal	Both	385	227	123	275
Fail#4	Both	385	227	118	280
Foolish	Both	385	221	121	283
Forbid	Negative	385	224	121	280
Fraught	Negative	385	223	122	280
Fruitless	Both	385	228	114	283
Guilt	Negative	385	225	121	279
Guise	Negative	385	224	121	280
Humiliate	Negative	385	228	114	283
Lament	Negative	385	220	122	283
Lazy	Both	385	219	123	283
Massacre	Negative	385	219	123	283
Mundane	Both	385	221	121	283
Negate	Negative	385	224	123	278
Plot#2	Negative	385	222	121	282
Rebel#1	Negative	385	221	121	283
Regrettable	Both	385	228	119	278
Ruthless	Negative	385	225	119	281
Sedentary	Negative	385	228	123	274
Shock#3	Negative	385	225	118	282
Shoot#3	Both	385	219	123	283
Simplistic	Negative	385	223	122	280
Storm#2	Negative	385	223	122	280
Temptation	Both	385	220	122	283
Tragic	Both	385	228	120	277
Unequal	Negative	385	222	123	280
Wear#4	Vice	385	224	123	278
Wilt	Negative	385	219	123	283
Ambiguous	Negative	386	223	122	281
Awkward	Both	386	226	119	281
Baffle	Negative	386	222	121	283
Belated	Negative	386	225	120	281
Calamity	Negative	386	222	123	281
Conflict#3	Negative	386	226	119	281

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Conspiracy	Negative	386	223	123	280
Contradiction	Negative	386	224	122	280
Darn	Negative	386	221	122	283
Deadlock	Negative	386	223	120	283
Demoralize	Negative	386	222	121	283
Disorganized	Both	386	228	123	275
Dreary	Both	386	221	123	282
Gloat	Negative	386	227	116	283
Grudge	Negative	386	220	123	283
Grumble	Negative	386	222	121	283
Hag	Negative	386	228	123	275
Humiliation	Negative	386	227	119	280
Insult	Negative	386	227	121	278
Matter#5	Both	386	226	117	283
Nullify	Negative	386	225	122	279
Particular#2	Both	386	225	120	281
Perilous	Both	386	222	121	283
Prisoner	Negative	386	226	123	277
Scoff	Negative	386	220	123	283
Scorch	Negative	386	223	123	280
Scream#1	Negative	386	220	123	283
Secrecy	Negative	386	228	116	282
Silly	Both	386	220	123	283
Slaughter	Negative	386	223	123	280
Squander	Negative	386	225	122	279
Study#3	Both	386	228	123	275
Thrash	Negative	386	221	122	283
Vie	Negative	386	226	123	277
Wound#5	Both	386	226	123	277
Wreck	Negative	386	222	122	282
Acrimonious	Negative	387	228	117	282
Antagonize	Negative	387	228	123	276
Ashamed	Negative	387	228	116	283
Bomb#2	Negative	387	223	121	283
Broke#5	Negative	387	228	116	283
Clutter	Both	387	227	123	277
Cold#2	Both	387	227	123	277
Confine#1	Negative	387	225	122	280
Cross#6	Negative	387	228	123	276
Dirt	Negative	387	224	123	280
Dread	Negative	387	225	123	279
Edge#4	Negative	387	222	122	283
Evil#1	Both	387	226	119	282
Feud	Negative	387	225	119	283
Forfeit	Negative	387	228	122	277
Founder#2	Negative	387	225	119	283
Frail	Vice	387	228	123	276
Frantically	Negative	387	222	122	283
Grief	Negative	387	227	117	283
Helpless	Negative	387	227	118	282
Infamous	Both	387	226	122	279
Insidious	Both	387	227	123	277
Mad#1	Negative	387	225	121	281
Mad#2	Both	387	225	123	279
Malady	Negative	387	228	123	276

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Merciless	Negative	387	221	123	283
Mishandle	Negative	387	228	120	279
Offender	Negative	387	226	122	279
Overpower	Negative	387	223	123	281
Powerless	Negative	387	226	119	282
Press#4	Negative	387	223	122	282
Resentment	Negative	387	228	116	283
Rogue	Negative	387	228	117	282
Shark	Negative	387	225	123	279
Snare	Negative	387	222	123	282
Snatch	Negative	387	223	121	283
Stab	Negative	387	223	122	282
Stern	Negative	387	224	122	281
Stink	Negative	387	222	123	282
Stubborn	Both	387	224	122	281
Thorny	Negative	387	222	123	282
Thud	Negative	387	221	123	283
Timid	Vice	387	222	122	283
Undone	Negative	387	222	122	283
Unreliable	Both	387	226	120	281
Uproar	Negative	387	227	121	279
Upset#3	Negative	387	223	122	282
Yelp	Negative	387	227	123	277
Absurd	Both	388	224	123	281
Adversity	Both	388	227	123	278
Aggressiveness	Both	388	226	122	280
Arduous	Both	388	226	122	280
Aversion	Negative	388	224	123	281
Bait	Negative	388	224	122	282
Bombard	Negative	388	224	122	282
Brazen	Negative	388	228	122	278
Chase#1	Negative	388	224	122	282
Choke#1	Negative	388	225	122	281
Clumsy	Both	388	226	120	282
Cold#3	Negative	388	222	123	283
Collision	Negative	388	227	123	278
Condemn#2	Negative	388	227	120	281
Conflict#4	Negative	388	227	123	278
Contradictory	Negative	388	224	122	282
Damn	Negative	388	222	123	283
Disappoint#4	Both	388	227	123	278
Discord	Both	388	227	123	278
Doomsday	Negative	388	226	120	282
Fed#2	Negative	388	224	121	283
Fix#5	Both	388	224	122	282
Fury	Negative	388	223	123	282
Grievance	Negative	388	228	120	280
Hazy	Negative	388	224	121	283
Horror	Negative	388	223	122	283
Hustle	Negative	388	225	121	282
Hysteria	Both	388	223	122	283
Impurity	Both	388	228	121	279
Inaccuracy	Negative	388	227	122	279
Malicious	Negative	388	228	123	277
Meek	Negative	388	223	123	282

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Mishap	Negative	388	224	122	282
Moan	Negative	388	222	123	283
Moody	Negative	388	225	123	280
Murder#1	Both	388	224	122	282
Predicament	Both	388	227	122	279
Preposterous	Both	388	223	123	282
Prohibition	Negative	388	228	123	277
Push#3	Negative	388	222	123	283
Recalcitrant	Both	388	227	123	278
Redundancy	Both	388	228	121	279
Regret#2	Negative	388	224	123	281
Rivalry	Negative	388	225	123	280
Rupture#2	Negative	388	225	122	281
Sabotage	Negative	388	225	121	282
Scapegoat	Negative	388	227	119	282
Scorn	Negative	388	223	122	283
Scrape	Negative	388	225	122	281
Shrivel	Negative	388	227	119	282
Shroud	Negative	388	226	120	282
Sick#2	Both	388	228	123	277
Sorry#2	Both	388	224	121	283
Stamp#2	Negative	388	225	123	280
Stormy	Negative	388	223	122	283
Strangle	Negative	388	226	123	279
Stunt	Both	388	224	123	281
Terror	Negative	388	227	119	282
Unimportant	Both	388	224	121	283
Unpopular	Both	388	226	123	279
Unseen	Negative	388	224	122	282
Unsteady	Negative	388	224	123	281
Warfare	Negative	388	224	123	281
Whimper	Negative	388	222	123	283
Worn#2	Negative	388	224	123	281
Annoy	Negative	389	225	121	283
Assail	Negative	389	226	120	283
Beat#4	Both	389	223	123	283
Bland	Negative	389	223	123	283
Break#4	Negative	389	228	121	280
Coercive	Negative	389	227	123	279
Confess	Negative	389	223	123	283
Confession	Negative	389	223	123	283
Congested	Negative	389	228	122	279
Cruel	Both	389	223	123	283
Curt	Negative	389	228	122	279
Deceive#1	Both	389	228	123	278
Deception	Both	389	228	120	281
Despair#1	Negative	389	224	123	282
Disapproval	Both	389	228	122	279
Disguise	Negative	389	226	122	281
Domination	Both	389	225	123	281
Fearsome	Negative	389	225	122	282
Frown#2	Negative	389	223	123	283
Fumble	Negative	389	223	123	283
Hopeless	Negative	389	225	123	281
Illegality	Both	389	228	123	278

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Incompatible	Both	389	228	122	279
Intolerable	Negative	389	228	123	278
Irk	Negative	389	225	122	282
Kick#3	Negative	389	224	122	283
Kill#2	Negative	389	227	122	280
Lost#4	Negative	389	225	123	281
Lying#2	Negative	389	228	123	278
Lying#3	Negative	389	228	122	279
Madness	Both	389	225	123	281
Mistake#3	Negative	389	228	119	282
Nix	Negative	389	228	121	280
Obstruct	Negative	389	228	123	278
Oppose#3	Negative	389	224	122	283
Outcry	Negative	389	226	123	280
Paranoid	Negative	389	228	123	278
Pointless	Negative	389	228	123	278
Prowl	Negative	389	224	122	283
Pry	Negative	389	227	119	283
Remorse	Negative	389	228	119	282
Reprehensible	Both	389	228	123	278
Resentful	Negative	389	228	118	283
Root#4	Negative	389	227	120	282
Ruin#1	Both	389	228	119	282
Sadness	Negative	389	228	123	278
Sequester	Negative	389	228	123	278
Serve#4	Negative	389	228	122	279
Sever	Negative	389	227	122	280
Shameful	Negative	389	227	123	279
Starve	Negative	389	227	122	280
Trivial	Negative	389	224	122	283
Trudge	Negative	389	223	123	283
Unnoticed	Negative	389	227	123	279
Vehement	Negative	389	227	122	280
Vexing	Negative	389	228	119	282
Wrath	Negative	389	224	122	283
Acrimony	Both	390	228	119	283
Addict	Negative	390	226	123	281
Agony	Negative	390	225	122	283
Alas	Negative	390	224	123	283
Alienate	Negative	390	227	122	281
Anguish	Negative	390	226	122	282
Antiquated	Negative	390	228	121	281
Audacity	Both	390	228	121	281
Besiege	Negative	390	225	123	282
Betray	Negative	390	227	120	283
Bit#2	Negative	390	225	123	282
Bogus	Both	390	227	121	282
Boredom	Both	390	224	123	283
Brag	Negative	390	226	121	283
Brittle	Negative	390	228	123	279
Cannon	Negative	390	226	123	281
Capricious	Both	390	225	123	282
Cheapen	Negative	390	224	123	283
Collide	Negative	390	225	123	282
Daze	Negative	390	225	122	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Decadent	Both	390	228	123	279
Desert#2	Negative	390	225	123	282
Detract	Negative	390	226	122	282
Devoid	Negative	390	227	122	281
Discredit	Negative	390	227	121	282
Dishonest	Negative	390	228	123	279
Dislike#2	Negative	390	226	123	281
Distracting	Negative	390	227	123	280
Divorce#1	Both	390	227	123	280
Divorce#2	Negative	390	225	122	283
Drown#2	Negative	390	226	122	282
Esoteric	Negative	390	224	123	283
Foe	Negative	390	226	122	282
Frigid	Negative	390	224	123	283
Front#3	Both	390	228	119	283
Germ	Negative	390	228	123	279
Grizzly	Negative	390	224	123	283
Growl	Negative	390	224	123	283
Haggard	Negative	390	228	119	283
Harass	Negative	390	228	123	279
Head#7	Vice	390	224	123	283
Heartless	Negative	390	228	123	279
Hindrance	Both	390	228	123	279
Horrify	Negative	390	226	123	281
Incompetence	Negative	390	228	119	283
Inequality	Negative	390	228	123	279
Inexplicable	Negative	390	228	123	279
Infest	Negative	390	227	123	280
Insecurity	Both	390	225	122	283
Intimidate	Negative	390	227	123	280
Invisible	Negative	390	227	122	281
Limit#4	Both	390	227	122	281
Load#3	Both	390	224	123	283
Misfortune	Both	390	226	122	282
Nonsense	Both	390	226	123	281
Phobia	Negative	390	224	123	283
Pinch#2	Negative	390	224	123	283
Qualm	Negative	390	224	123	283
Quarrel#2	Negative	390	225	122	283
Regress	Negative	390	225	123	282
Repudiate	Negative	390	227	121	282
Rot	Negative	390	227	123	280
Screech#1	Negative	390	224	123	283
Shadow#2	Negative	390	226	123	281
Shirk	Negative	390	226	123	281
Shortsighted	Negative	390	227	122	281
Sick#3	Both	390	227	123	280
Slug	Both	390	224	123	283
Smuggle	Negative	390	227	122	281
Subvert	Negative	390	228	123	279
Taboo	Negative	390	228	123	279
Tease	Negative	390	224	123	283
Terrify	Negative	390	224	123	283
Twitch	Negative	390	227	123	280
Unchecked	Vice	390	226	123	281

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Unfamiliar	Both	390	225	122	283
Uninformed	Negative	390	228	122	280
Unqualified	Both	390	228	122	280
Useless	Both	390	226	122	282
Adulterate	Negative	391	228	123	280
Ambiguity	Negative	391	226	123	282
Ambivalent	Negative	391	228	122	281
Animosity	Both	391	227	121	283
Antagonistic	Negative	391	228	122	281
Arrogant	Both	391	226	122	283
Assault#2	Negative	391	226	123	282
Bizarre	Both	391	226	123	282
Blah	Negative	391	225	123	283
Bloodshed	Negative	391	225	123	283
Blunder	Negative	391	226	122	283
Boisterous	Negative	391	225	123	283
Bore#5	Both	391	225	123	283
Burglary	Both	391	228	121	282
Chastise	Negative	391	227	123	281
Chore	Negative	391	227	123	281
Cocky	Negative	391	227	121	283
Complicity	Both	391	227	121	283
Curse#1	Negative	391	227	123	281
Damned	Both	391	225	123	283
Defiance	Both	391	226	122	283
Defiant	Both	391	227	123	281
Delinquent	Both	391	225	123	283
Delirium	Both	391	228	123	280
Deride	Negative	391	226	123	282
Despair#2	Negative	391	226	123	282
Detain	Negative	391	226	122	283
Discouragement	Both	391	225	123	283
Disgust#4	Negative	391	226	123	282
Displeasure	Negative	391	226	123	282
Distrustful	Negative	391	226	123	282
Dumb	Negative	391	225	123	283
Exaggeration	Negative	391	227	122	282
Exhaustion	Negative	391	226	123	282
Fabricate	Negative	391	228	123	280
Ferocity	Negative	391	225	123	283
Fiasco	Negative	391	227	121	283
Fist	Negative	391	227	122	282
Forbidden	Negative	391	228	122	281
Fretful	Negative	391	225	123	283
Hassle	Negative	391	228	123	280
Horde	Negative	391	225	123	283
Hunt#3	Negative	391	226	122	283
Ignorance	Both	391	227	121	283
Immature	Negative	391	228	123	280
Incompetent	Negative	391	228	120	283
Inconvenient	Negative	391	228	121	282
Inferiority	Negative	391	228	123	280
Inundated	Negative	391	225	123	283
Jail#2	Negative	391	228	120	283
Jumpy	Negative	391	226	122	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Meddle	Negative	391	227	123	281
Mourn#2	Negative	391	225	123	283
Muddy	Negative	391	226	122	283
Obliterate	Negative	391	226	122	283
Overflow	Negative	391	228	121	282
Pandemonium	Negative	391	225	123	283
Pick#3	Negative	391	225	123	283
Poisonous	Negative	391	225	123	283
Primitive	Vice	391	228	123	280
Quarrel#1	Negative	391	227	121	283
Ramble	Negative	391	225	123	283
Renounce	Negative	391	228	122	281
Resent	Negative	391	227	121	283
Retard	Negative	391	225	123	283
Revenge	Negative	391	226	122	283
Rip#1	Negative	391	228	122	281
Rotten	Both	391	226	123	282
Service#3	Negative	391	226	123	282
Shred	Negative	391	226	123	282
Sinister	Both	391	226	123	282
Skirmish	Negative	391	225	123	283
Slight#4	Negative	391	226	122	283
Snore	Negative	391	228	123	280
Somber	Negative	391	228	120	283
Tatter	Negative	391	227	121	283
Tempest	Negative	391	225	123	283
Tire#3	Negative	391	226	123	282
Tire#4	Negative	391	225	123	283
Unnatural	Negative	391	227	123	281
Upset#4	Negative	391	226	122	283
Vex	Negative	391	225	123	283
Whip#1	Negative	391	226	122	283
Wrought	Both	391	226	123	282
Yearn	Negative	391	225	123	283
Abdicate	Negative	392	227	122	283
Abhor	Negative	392	226	123	283
Abolish	Negative	392	226	123	283
Abyss	Negative	392	227	123	282
Aghast	Both	392	226	123	283
Aloof	Negative	392	226	123	283
Amiss	Both	392	227	123	282
Annihilate	Negative	392	226	123	283
Anxiousness	Negative	392	227	123	282
Apathetic	Negative	392	226	123	283
Assassinate	Negative	392	226	123	283
Astray	Negative	392	228	121	283
Avenge	Both	392	226	123	283
Babble	Negative	392	226	123	283
Bane	Both	392	227	123	282
Barren	Negative	392	227	122	283
Bewilderment	Negative	392	226	123	283
Bite#3	Negative	392	227	123	282
Blind#3	Both	392	227	122	283
Box#2	Negative	392	227	123	282
Bum	Negative	392	226	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Careless	Both	392	227	123	282
Cave#2	Negative	392	227	122	283
Censorship	Negative	392	228	123	281
Chide	Negative	392	228	122	282
Condemnation	Both	392	228	123	281
Condescending	Both	392	227	122	283
Consternation	Both	392	226	123	283
Corrosion	Negative	392	228	122	282
Coward	Negative	392	227	122	283
Cram	Negative	392	226	123	283
Cranky	Negative	392	227	123	282
Cringe	Negative	392	226	123	283
Crook	Both	392	228	121	283
Crumple	Negative	392	226	123	283
Dawdle	Negative	392	228	121	283
Deaf	Both	392	227	123	282
Deafness	Both	392	228	123	281
Depose	Negative	392	228	123	281
Derision	Both	392	227	122	283
Derisive	Both	392	226	123	283
Disapprove	Negative	392	228	121	283
Disarm	Negative	392	227	123	282
Disavow	Negative	392	226	123	283
Disavowal	Negative	392	228	121	283
Disconcerted	Negative	392	226	123	283
Discontent	Negative	392	227	123	282
Discordant	Negative	392	228	123	281
Disgrace	Both	392	227	123	282
Dislike#1	Negative	392	227	123	282
Divorce#3	Negative	392	228	122	282
Drab	Both	392	226	123	283
Drown#1	Negative	392	227	123	282
Entangle	Negative	392	227	122	283
Exasperate	Negative	392	226	123	283
Exception#2	Negative	392	226	123	283
Expedient	Negative	392	228	122	282
Extravagant	Both	392	227	122	283
Fat#3	Negative	392	226	123	283
Ferocious	Negative	392	226	123	283
Flimsy	Negative	392	228	123	281
Forsake	Negative	392	227	122	283
Freak	Negative	392	226	123	283
Frightful	Negative	392	226	123	283
Gall	Negative	392	228	123	281
Grave#4	Both	392	227	123	282
Grotesque	Both	392	228	121	283
Hang#6	Negative	392	226	123	283
Haphazard	Negative	392	227	123	282
Hypocrisy	Both	392	228	123	281
Imperfect	Both	392	226	123	283
Implore	Negative	392	228	123	281
Imprisonment	Negative	392	227	123	282
Incapable	Both	392	228	122	282
Indeterminate	Negative	392	227	123	282
Inexact	Both	392	227	123	282

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Infraction	Negative	392	227	123	282
Insane	Both	392	226	123	283
Insinuate	Negative	392	228	122	282
Jerk#1	Negative	392	227	123	282
Laugh#3	Negative	392	226	123	283
Lifeless	Negative	392	226	123	283
Lowly	Negative	392	227	122	283
Ludicrous	Both	392	227	122	283
Lunatic	Both	392	226	123	283
Mess#2	Negative	392	228	121	283
Mind#9	Negative	392	228	123	281
Mistrust	Negative	392	227	123	282
Misunderstand	Negative	392	228	123	281
Mock	Negative	392	228	123	281
Motley	Negative	392	227	122	283
Mutter	Negative	392	226	123	283
Offend	Negative	392	228	121	283
Ordeal	Both	392	228	122	282
Outburst	Negative	392	226	123	283
Outcast	Negative	392	226	123	283
Pathetic	Negative	392	226	123	283
Perish	Negative	392	226	123	283
Pitiful	Negative	392	226	123	283
Pollute	Negative	392	227	122	283
Prejudicial	Negative	392	228	123	281
Procrastination	Negative	392	226	123	283
Provocation	Negative	392	226	123	283
Puzzlement	Negative	392	226	123	283
Reactionary	Negative	392	227	123	282
Rebuke	Negative	392	228	123	281
Rebut	Negative	392	227	122	283
Reproach	Negative	392	228	122	282
Revolt	Negative	392	226	123	283
Rubbish	Negative	392	228	122	282
Ruin#4	Negative	392	228	121	283
Ruinous	Both	392	228	121	283
Sarcastic	Negative	392	226	123	283
Screw#2	Negative	392	227	123	282
Segregation	Negative	392	228	122	282
Senile	Negative	392	228	123	281
Sentence#2	Negative	392	227	123	282
Shameless	Both	392	228	123	281
Shock#5	Negative	392	228	123	281
Siege	Negative	392	227	123	282
Sleepless	Negative	392	227	122	283
Smother	Negative	392	227	123	282
Sneer	Negative	392	227	122	283
Sob	Negative	392	226	123	283
Sprain	Negative	392	228	123	281
Steal#2	Both	392	228	123	281
Sunburn	Vice	392	227	123	282
Swore	Negative	392	227	123	282
Tedious	Both	392	226	123	283
Thief	Negative	392	228	123	281
Throw#3	Negative	392	227	122	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Torment	Negative	392	226	123	283
Touchy	Both	392	226	123	283
Trick#2	Negative	392	227	123	282
Ultimatum	Negative	392	226	123	283
Unfit	Both	392	228	123	281
Unmoved	Negative	392	226	123	283
Unprepared	Vice	392	228	121	283
Unreal	Vice	392	227	122	283
Unsatisfactory	Both	392	228	123	281
Untimely	Negative	392	227	123	282
Unwise	Both	392	226	123	283
Uprising	Negative	392	226	123	283
Vanity	Both	392	228	122	282
Venom	Both	392	228	123	281
Wild#2	Negative	392	227	122	283
Abject	Both	393	228	122	283
Abrasive	Both	393	228	123	282
Absurdity	Both	393	227	123	283
Admonish	Negative	393	228	123	282
Adversary	Negative	393	228	122	283
Affliction	Both	393	228	123	282
Aggravation	Both	393	228	123	282
Alien	Negative	393	227	123	283
Amputate	Negative	393	228	123	282
Annihilation	Both	393	227	123	283
Anomalous	Negative	393	227	123	283
Antipathy	Both	393	227	123	283
Appall#2	Negative	393	228	123	282
Assassin	Negative	393	227	123	283
Atrocious	Both	393	227	123	283
Audacious	Both	393	228	122	283
Austere	Negative	393	228	122	283
Bashful	Negative	393	228	122	283
Belittle	Negative	393	228	122	283
Belligerent	Negative	393	227	123	283
Betrayal	Negative	393	228	122	283
Bombardment	Negative	393	227	123	283
Bondage	Both	393	227	123	283
Bore#1	Both	393	227	123	283
Bother#3	Both	393	228	122	283
Brat	Both	393	228	122	283
Bravado	Negative	393	228	122	283
Brood	Negative	393	227	123	283
Brusque	Negative	393	227	123	283
Brutish	Both	393	227	123	283
Bug#2	Negative	393	227	123	283
Bungle	Negative	393	228	122	283
Callous	Negative	393	228	123	282
Cataclysm	Negative	393	227	123	283
Chafe	Negative	393	227	123	283
Clatter	Negative	393	227	123	283
Coarse	Negative	393	227	123	283
Coerce	Negative	393	228	123	282
Commotion	Both	393	227	123	283
Condemn#1	Negative	393	228	123	282

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Confinement	Both	393	227	123	283
Convict#1	Negative	393	227	123	283
Convict#2	Negative	393	228	122	283
Corrosive	Negative	393	228	123	282
Covert	Both	393	228	122	283
Craziness	Both	393	227	123	283
Culpable	Negative	393	228	122	283
Curse#2	Negative	393	227	123	283
Cursory	Both	393	228	123	282
Cynicism	Both	393	227	123	283
Debatable	Negative	393	228	123	282
Deceit	Both	393	227	123	283
Degenerate	Both	393	227	123	283
Dejected	Negative	393	227	123	283
Demon	Both	393	227	123	283
Deplore	Negative	393	227	123	283
Desert#4	Negative	393	227	123	283
Despicable	Both	393	228	123	282
Disadvantageous	Both	393	227	123	283
Disgust#1	Negative	393	227	123	283
Displease	Negative	393	227	123	283
Disprove	Negative	393	227	123	283
Dissatisfy	Negative	393	227	123	283
Dope	Both	393	227	123	283
Doubt#4	Negative	393	227	123	283
Downcast	Negative	393	228	123	282
Drowsy	Negative	393	227	123	283
Drunk#1	Both	393	227	123	283
Eccentric	Negative	393	228	122	283
Encroach	Negative	393	228	123	282
Enslave	Negative	393	227	123	283
Envious	Both	393	228	123	282
Epithet	Negative	393	227	123	283
Even#3	Negative	393	228	123	282
Expel	Negative	393	228	123	282
Extinct	Negative	393	227	123	283
Falsehood	Negative	393	228	123	282
Famine	Negative	393	227	123	283
Fanatical	Both	393	228	122	283
Farce	Negative	393	228	123	282
Feint	Negative	393	227	123	283
Fidget	Negative	393	228	122	283
Flagrant	Negative	393	228	123	282
Floor#5	Negative	393	227	123	283
Foible	Negative	393	227	123	283
Force#5	Negative	393	227	123	283
Foreboding	Negative	393	227	123	283
Forgetfulness	Both	393	228	123	282
Forlorn	Negative	393	227	123	283
Fuss	Negative	393	227	123	283
Gaudy	Negative	393	227	123	283
Ghetto	Negative	393	227	123	283
Gratuitous	Negative	393	228	123	282
Grieve	Negative	393	228	123	282
Guerrilla	Negative	393	227	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Gullible	Negative	393	227	123	283
Hate#3	Negative	393	227	123	283
Hater	Negative	393	227	123	283
Hatred	Negative	393	228	123	282
Hazard#2	Negative	393	227	123	283
Hit#4	Negative	393	227	123	283
Homely	Both	393	228	122	283
Horrid	Both	393	227	123	283
Idiot	Both	393	227	123	283
Illogical	Negative	393	227	123	283
Immobility	Negative	393	228	123	282
Impatience	Negative	393	227	123	283
Impersonal	Both	393	228	123	282
Impulsive	Both	393	228	123	282
Incessant	Negative	393	228	123	282
Incompatibility	Both	393	228	123	282
Indecisiveness	Vice	393	227	123	283
Indignation	Negative	393	228	122	283
Indulge	Negative	393	228	123	282
Infuriate	Negative	393	228	122	283
Insecure	Both	393	228	123	282
Intrude	Negative	393	228	122	283
Inundate	Negative	393	227	123	283
Iron#3	Negative	393	227	123	283
Jeer	Negative	393	228	122	283
Jerk#2	Negative	393	227	123	283
Kidnap	Negative	393	227	123	283
Lame	Negative	393	227	123	283
Lazily	Negative	393	227	123	283
Loneliness	Negative	393	228	123	282
Loner	Negative	393	228	122	283
Malice	Negative	393	228	123	282
Mangle	Negative	393	228	122	283
Misinformed	Negative	393	228	123	282
Monotonous	Both	393	227	123	283
Monstrous	Both	393	228	122	283
Murder#2	Negative	393	227	123	283
Needle#2	Negative	393	228	123	282
Nuisance	Both	393	228	123	282
Oblique	Negative	393	228	123	282
Obstinate	Negative	393	227	123	283
Oppressive	Negative	393	227	123	283
Outlaw	Negative	393	227	123	283
Overthrow	Negative	393	227	123	283
Partition	Negative	393	228	123	282
Perplexity	Negative	393	227	123	283
Persecution	Negative	393	228	123	282
Petty	Both	393	228	122	283
Pretend	Negative	393	228	123	282
Pretense	Negative	393	228	123	282
Procrastinate	Negative	393	228	123	282
Queer	Both	393	228	122	283
Quitter	Negative	393	228	123	282
Rebellion	Negative	393	227	123	283
Recoil	Negative	393	227	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Refuse#2	Vice	393	228	123	282
Robber	Negative	393	227	123	283
Roughness	Both	393	228	123	282
Ruin#3	Both	393	228	122	283
Rusty	Negative	393	227	123	283
Scandalous	Negative	393	228	122	283
Scheme#2	Negative	393	228	122	283
Scuffle	Negative	393	227	123	283
Secede	Negative	393	227	123	283
Seethe	Negative	393	227	123	283
Senseless	Negative	393	228	122	283
Shabby	Negative	393	227	123	283
Shady	Both	393	228	123	282
Shoddy	Negative	393	227	123	283
Sick#4	Both	393	227	123	283
Snarl	Negative	393	228	123	282
Sorrow	Negative	393	228	122	283
Spank	Negative	393	227	123	283
Spear#2	Negative	393	228	123	282
Stale	Both	393	228	123	282
Starvation	Negative	393	228	123	282
Stupid#2	Both	393	228	122	283
Stupidity	Both	393	228	122	283
Subjugate	Negative	393	228	123	282
Subversion	Negative	393	228	122	283
Suffocate	Negative	393	228	122	283
Superstition	Both	393	227	123	283
Superstitious	Both	393	227	123	283
Tantrum	Negative	393	227	123	283
Tardy	Negative	393	227	123	283
Taunt#1	Negative	393	227	123	283
Taunt#2	Negative	393	227	123	283
Terrorize	Negative	393	227	123	283
Timidity	Both	393	227	123	283
Tire#2	Negative	393	227	123	283
Tired#4	Negative	393	227	123	283
Toil	Negative	393	227	123	283
Torturous	Negative	393	227	123	283
Tramp	Negative	393	228	122	283
Unbearable	Negative	393	228	123	282
Undid	Negative	393	227	123	283
Unhappiness	Negative	393	228	122	283
Unlucky	Both	393	227	123	283
Unmitigated	Vice	393	228	122	283
Unobserved	Negative	393	228	123	282
Unscrupulous	Both	393	228	123	282
Unspeakable	Negative	393	227	123	283
Unsteadiness	Negative	393	227	123	283
Unworthy	Both	393	227	123	283
Usurp	Negative	393	228	122	283
Villain	Negative	393	227	123	283
Wail	Negative	393	227	123	283
Weariness	Negative	393	227	123	283
Wearisome	Negative	393	228	122	283
Weird	Both	393	227	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Wicked	Negative	393	227	123	283
Wild#3	Negative	393	227	123	283
Worn#3	Both	393	228	123	282
Abominable	Both	394	228	123	283
Abscond	Negative	394	228	123	283
Absent-Minded	Negative	394	228	123	283
Absentee	Both	394	228	123	283
Accost	Negative	394	228	123	283
Accursed	Both	394	228	123	283
Admonition	Negative	394	228	123	283
Adulteration	Both	394	228	123	283
Adultery	Both	394	228	123	283
Affectation	Both	394	228	123	283
Afraid#2	Negative	394	228	123	283
Aggressor	Negative	394	228	123	283
Agitator	Negative	394	228	123	283
Alibi	Negative	394	228	123	283
Alienation	Both	394	228	123	283
Altercation	Negative	394	228	123	283
Ambush#1	Negative	394	228	123	283
Ambush#2	Negative	394	228	123	283
Anarchist	Negative	394	228	123	283
Anarchy	Negative	394	228	123	283
Annoyance	Negative	394	228	123	283
Anti-Social	Both	394	228	123	283
Appall#1	Both	394	228	123	283
Assailant	Negative	394	228	123	283
Asunder	Negative	394	228	123	283
Autocrat	Negative	394	228	123	283
Autocratic	Negative	394	228	123	283
Avarice	Both	394	228	123	283
Avaricious	Both	394	228	123	283
Awkwardness	Both	394	228	123	283
Backwardness	Both	394	228	123	283
Bafflement	Negative	394	228	123	283
Banal	Both	394	228	123	283
Bandit	Negative	394	228	123	283
Banish#1	Negative	394	228	123	283
Banish#2	Negative	394	228	123	283
Banishment	Negative	394	228	123	283
Barbarian	Both	394	228	123	283
Barbarous	Both	394	228	123	283
Bastard	Negative	394	228	123	283
Beastly	Both	394	228	123	283
Beggar	Negative	394	228	123	283
Behead	Negative	394	228	123	283
Belt#2	Negative	394	228	123	283
Bereave	Negative	394	228	123	283
Bereavement	Both	394	228	123	283
Bereft	Negative	394	228	123	283
Berserk	Negative	394	228	123	283
Beseech	Negative	394	228	123	283
Beside#3	Vice	394	228	123	283
Bitchy	Negative	394	228	123	283
Bitterness	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Blackmail	Negative	394	228	123	283
Blemish	Both	394	228	123	283
Blockhead	Both	394	228	123	283
Bloodthirsty	Negative	394	228	123	283
Blurt	Negative	394	228	123	283
Board#7	Both	394	228	123	283
Boastful	Both	394	228	123	283
Bow#3	Vice	394	228	123	283
Brandish	Negative	394	228	123	283
Brawl	Negative	394	228	123	283
Bristle	Negative	394	228	123	283
Broken-Hearted	Negative	394	228	123	283
Brutality	Both	394	228	123	283
Brute	Both	394	228	123	283
Burglar	Both	394	228	123	283
Burn#4	Negative	394	228	123	283
Busybody	Both	394	228	123	283
Butchery	Both	394	228	123	283
Cannibal	Both	394	228	123	283
Capital#2	Both	394	228	123	283
Capsize	Negative	394	228	123	283
Carelessness	Both	394	228	123	283
Catch#3	Both	394	228	123	283
Censure	Negative	394	228	123	283
Cheater	Both	394	228	123	283
Childish	Both	394	228	123	283
Circle#3	Both	394	228	123	283
Clamorous	Negative	394	228	123	283
Clique	Negative	394	228	123	283
Club#2	Negative	394	228	123	283
Club#3	Negative	394	228	123	283
Coarseness	Both	394	228	123	283
Cockiness	Both	394	228	123	283
Coercion	Both	394	228	123	283
Cold#6	Negative	394	228	123	283
Collusion	Negative	394	228	123	283
Combatant	Negative	394	228	123	283
Commiseration	Both	394	228	123	283
Commoner	Negative	394	228	123	283
Compulsion	Negative	394	228	123	283
Conceit	Both	394	228	123	283
Condescension	Both	394	228	123	283
Confiscate	Negative	394	228	123	283
Confiscation	Both	394	228	123	283
Conspirator	Negative	394	228	123	283
Consumptive	Both	394	228	123	283
Contemptible	Both	394	228	123	283
Contemptuous	Both	394	228	123	283
Coolness	Negative	394	228	123	283
Corrode	Negative	394	228	123	283
Costliness	Negative	394	228	123	283
Counteraction	Negative	394	228	123	283
Crafty	Both	394	228	123	283
Crass	Both	394	228	123	283
Crawl#2	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Credulous	Negative	394	228	123	283
Croak	Negative	394	228	123	283
Crooked	Both	394	228	123	283
Cross#3	Negative	394	228	123	283
Cross#7	Negative	394	228	123	283
Cruelty	Both	394	228	123	283
Damnably	Both	394	228	123	283
Deadweight	Both	394	228	123	283
Decadence	Both	394	228	123	283
Decay#2	Negative	394	228	123	283
Decease	Negative	394	228	123	283
Deceitful	Both	394	228	123	283
Decompose	Negative	394	228	123	283
Defame	Negative	394	228	123	283
Defile	Negative	394	228	123	283
Demand#4	Vice	394	228	123	283
Demean	Negative	394	228	123	283
Demolish	Negative	394	228	123	283
Deplorable	Both	394	228	123	283
Depraved	Both	394	228	123	283
Derogatory	Negative	394	228	123	283
Desertion	Both	394	228	123	283
Desolate	Both	394	228	123	283
Desolation	Both	394	228	123	283
Despise#1	Both	394	228	123	283
Despise#2	Negative	394	228	123	283
Destitute	Negative	394	228	123	283
Detest	Negative	394	228	123	283
Detestable	Both	394	228	123	283
Devilish	Negative	394	228	123	283
Devious	Negative	394	228	123	283
Diabolic	Negative	394	228	123	283
Diabolical	Negative	394	228	123	283
Dictatorial	Negative	394	228	123	283
Disagreeable	Negative	394	228	123	283
Discrepant	Negative	394	228	123	283
Disgust#2	Negative	394	228	123	283
Disgust#3	Negative	394	228	123	283
Disgust#5	Negative	394	228	123	283
Dishonor	Negative	394	228	123	283
Disobedience	Both	394	228	123	283
Disobedient	Both	394	228	123	283
Dispensability	Negative	394	228	123	283
Disputable	Both	394	228	123	283
Dissention	Both	394	228	123	283
Dominate#3	Negative	394	228	123	283
Double#3	Negative	394	228	123	283
Double#4	Both	394	228	123	283
Downhearted	Negative	394	228	123	283
Drunk#2	Negative	394	228	123	283
Drunkard	Both	394	228	123	283
Drunken	Both	394	228	123	283
Dunce	Both	394	228	123	283
Dungeon	Negative	394	228	123	283
Eccentricity	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Egotistical	Both	394	228	123	283
Encroachment	Negative	394	228	123	283
Enrage	Negative	394	228	123	283
Entanglement	Negative	394	228	123	283
Entreat	Negative	394	228	123	283
Estranged	Negative	394	228	123	283
Evasion	Negative	394	228	123	283
Evict	Negative	394	228	123	283
Evil#2	Both	394	228	123	283
Exasperation	Negative	394	228	123	283
Excommunication	Negative	394	228	123	283
Exile	Negative	394	228	123	283
Exterminate	Negative	394	228	123	283
Extermination	Negative	394	228	123	283
Fail#3	Negative	394	228	123	283
Fallacy	Negative	394	228	123	283
Famished	Negative	394	228	123	283
Fanatic	Both	394	228	123	283
Fascist	Negative	394	228	123	283
Fatalistic	Negative	394	228	123	283
Feign	Negative	394	228	123	283
Feudal	Negative	394	228	123	283
Fiend	Negative	394	228	123	283
Filth	Both	394	228	123	283
Filthy	Both	394	228	123	283
Foolishness	Both	394	228	123	283
Foreign#2	Both	394	228	123	283
Fugitive	Negative	394	228	123	283
Fun#2	Negative	394	228	123	283
Fussy	Both	394	228	123	283
Gash	Negative	394	228	123	283
Ghastly	Both	394	228	123	283
Goddamn	Negative	394	228	123	283
Grave#3	Both	394	228	123	283
Grave#5	Both	394	228	123	283
Gruff	Negative	394	228	123	283
Gunmen	Negative	394	228	123	283
Hackney	Negative	394	228	123	283
Hand#11	Vice	394	228	123	283
Hand#6	Both	394	228	123	283
Hang#3	Negative	394	228	123	283
Hapless	Both	394	228	123	283
Hard#5	Negative	394	228	123	283
Haughty	Negative	394	228	123	283
Haziness	Negative	394	228	123	283
Hedonistic	Both	394	228	123	283
Heedless	Negative	394	228	123	283
Heinous	Negative	394	228	123	283
Help#3	Negative	394	228	123	283
Helplessness	Both	394	228	123	283
Hideous	Both	394	228	123	283
Hole#2	Both	394	228	123	283
Hole#3	Both	394	228	123	283
Horn#3	Both	394	228	123	283
Hot#6	Both	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Hung#3	Negative	394	228	123	283
Hurtful	Both	394	228	123	283
Hustler	Negative	394	228	123	283
Hypocrite	Both	394	228	123	283
Hysterical	Negative	394	228	123	283
Idiotic	Both	394	228	123	283
Idleness	Both	394	228	123	283
Ignoble	Both	394	228	123	283
Ignorant	Both	394	228	123	283
Ill#4	Negative	394	228	123	283
Illiterate	Negative	394	228	123	283
Immoral	Both	394	228	123	283
Immorality	Both	394	228	123	283
Immovable	Negative	394	228	123	283
Impetuous	Both	394	228	123	283
Imprecision	Negative	394	228	123	283
Imprison	Negative	394	228	123	283
Impure	Both	394	228	123	283
Inane	Both	394	228	123	283
Indecent	Both	394	228	123	283
Indeterminable	Negative	394	228	123	283
Ineffectiveness	Negative	394	228	123	283
Ineffectual	Negative	394	228	123	283
Ineffectualness	Negative	394	228	123	283
Ingratitude	Both	394	228	123	283
Inhumane	Both	394	228	123	283
Injurious	Both	394	228	123	283
Insensible	Negative	394	228	123	283
Insolence	Negative	394	228	123	283
Insolent	Negative	394	228	123	283
Instable	Both	394	228	123	283
Intoxicate	Negative	394	228	123	283
Intruder	Negative	394	228	123	283
Intrusion	Negative	394	228	123	283
Iron#5	Negative	394	228	123	283
Kick#4	Negative	394	228	123	283
Knife#2	Negative	394	228	123	283
Lamentable	Negative	394	228	123	283
Lawless	Negative	394	228	123	283
Let#3	Negative	394	228	123	283
Liar	Negative	394	228	123	283
Lone#1	Negative	394	228	123	283
Lonesome	Negative	394	228	123	283
Loveless	Both	394	228	123	283
Madman	Both	394	228	123	283
Maladjusted	Both	394	228	123	283
Maladjustment	Both	394	228	123	283
Manner#4	Vice	394	228	123	283
Manslaughter	Both	394	228	123	283
Melancholy	Negative	394	228	123	283
Melodramatic	Negative	394	228	123	283
Menace	Both	394	228	123	283
Menial	Negative	394	228	123	283
Misbehave	Negative	394	228	123	283
Misbehavior	Both	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Mischief	Negative	394	228	123	283
Mischievous	Negative	394	228	123	283
Miser	Negative	394	228	123	283
Misinform	Negative	394	228	123	283
Mistake#2	Both	394	228	123	283
Mix#3	Both	394	228	123	283
Mockery	Negative	394	228	123	283
Molest	Negative	394	228	123	283
Monotony	Both	394	228	123	283
Mortify	Negative	394	228	123	283
Motionless	Negative	394	228	123	283
Mourn#1	Negative	394	228	123	283
Mourner	Negative	394	228	123	283
Mumble	Negative	394	228	123	283
Murderous	Negative	394	228	123	283
Myself#3	Negative	394	228	123	283
Naughty	Negative	394	228	123	283
Nebulous	Negative	394	228	123	283
Negation	Both	394	228	123	283
Neurotic	Negative	394	228	123	283
Nonchalant	Negative	394	228	123	283
Nosey	Negative	394	228	123	283
Nullification	Negative	394	228	123	283
Obnoxious	Both	394	228	123	283
Oddity	Negative	394	228	123	283
Opinionated	Negative	394	228	123	283
Oppress	Negative	394	228	123	283
Order#8	Negative	394	228	123	283
Ostracize	Negative	394	228	123	283
Overbearing	Negative	394	228	123	283
Overworked	Negative	394	228	123	283
Pass#10	Negative	394	228	123	283
Pass#13	Negative	394	228	123	283
Pass#8	Negative	394	228	123	283
Patronize	Negative	394	228	123	283
Persecute	Negative	394	228	123	283
Perturb	Negative	394	228	123	283
Pervert	Both	394	228	123	283
Piece#4	Negative	394	228	123	283
Pitiless	Negative	394	228	123	283
Pompous	Negative	394	228	123	283
Pout	Negative	394	228	123	283
Presumptuous	Both	394	228	123	283
Pretentious	Both	394	228	123	283
Prison#2	Negative	394	228	123	283
Propaganda	Negative	394	228	123	283
Quarrelsome	Negative	394	228	123	283
Quibble	Negative	394	228	123	283
Race#4	Negative	394	228	123	283
Raise#5	Negative	394	228	123	283
Rascal	Both	394	228	123	283
Rebel#2	Negative	394	228	123	283
Rebellious	Negative	394	228	123	283
Recklessness	Both	394	228	123	283
Renunciation	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Repress	Negative	394	228	123	283
Repugnant	Both	394	228	123	283
Repulse	Negative	394	228	123	283
Repulsive	Both	394	228	123	283
Ride#3	Negative	394	228	123	283
Ridicule#1	Negative	394	228	123	283
Ridicule#2	Negative	394	228	123	283
Robbery	Negative	394	228	123	283
Roundabout	Both	394	228	123	283
Rue	Negative	394	228	123	283
Ruffian	Negative	394	228	123	283
Rumple	Negative	394	228	123	283
Run#4	Negative	394	228	123	283
Run#5	Negative	394	228	123	283
Ruthlessness	Negative	394	228	123	283
Sarcasm	Negative	394	228	123	283
Scald	Negative	394	228	123	283
Scold#1	Negative	394	228	123	283
Scold#2	Negative	394	228	123	283
Scornful	Negative	394	228	123	283
Scoundrel	Both	394	228	123	283
Scowl	Negative	394	228	123	283
Screech#2	Negative	394	228	123	283
Scum	Both	394	228	123	283
Secession	Negative	394	228	123	283
Selfish	Both	394	228	123	283
Selfishness	Negative	394	228	123	283
Serve#3	Vice	394	228	123	283
Servitude	Negative	394	228	123	283
Shaggy	Both	394	228	123	283
Shell#2	Negative	394	228	123	283
Shipwreck	Negative	394	228	123	283
Shoot#6	Negative	394	228	123	283
Show#4	Negative	394	228	123	283
Shrew	Negative	394	228	123	283
Shriek	Negative	394	228	123	283
Shrill	Negative	394	228	123	283
Shyness	Negative	394	228	123	283
Sin#2	Both	394	228	123	283
Sinful	Negative	394	228	123	283
Skulk	Negative	394	228	123	283
Slander	Negative	394	228	123	283
Slanderer	Negative	394	228	123	283
Slandorous	Negative	394	228	123	283
Slayer	Negative	394	228	123	283
Sleazy	Negative	394	228	123	283
Slight#5	Negative	394	228	123	283
Slime	Negative	394	228	123	283
Sloth	Both	394	228	123	283
Slothful	Both	394	228	123	283
Sly	Both	394	228	123	283
Smash#2	Negative	394	228	123	283
Sorrowful	Negative	394	228	123	283
Spinster	Negative	394	228	123	283
Spite#2	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Spiteful	Negative	394	228	123	283
Split#3	Negative	394	228	123	283
Spot#5	Both	394	228	123	283
Stammer	Negative	394	228	123	283
Stick#6	Both	394	228	123	283
Stolen#2	Negative	394	228	123	283
Straggle	Negative	394	228	123	283
Straggler	Negative	394	228	123	283
Strike#5	Both	394	228	123	283
Stubbornness	Negative	394	228	123	283
Stupid#3	Both	394	228	123	283
Stupor	Negative	394	228	123	283
Subjection	Negative	394	228	123	283
Subjugation	Negative	394	228	123	283
Submissive	Negative	394	228	123	283
Subservience	Negative	394	228	123	283
Sullen	Negative	394	228	123	283
Sunder	Negative	394	228	123	283
Superficiality	Both	394	228	123	283
Superfluous	Negative	394	228	123	283
Swear	Negative	394	228	123	283
Tear#3	Negative	394	228	123	283
Thirsty	Negative	394	228	123	283
Thoughtless	Negative	394	228	123	283
Throw#2	Negative	394	228	123	283
Tired#3	Negative	394	228	123	283
Tiresome	Both	394	228	123	283
Too#3	Both	394	228	123	283
Touch#7	Vice	394	228	123	283
Touch#9	Vice	394	228	123	283
Traitor	Negative	394	228	123	283
Transgress	Negative	394	228	123	283
Treachery	Negative	394	228	123	283
Treason	Negative	394	228	123	283
Treasonous	Negative	394	228	123	283
Trespass	Negative	394	228	123	283
Truant	Negative	394	228	123	283
Try#4	Negative	394	228	123	283
Try#5	Negative	394	228	123	283
Turn#5	Negative	394	228	123	283
Tyranny	Both	394	228	123	283
Unaccustomed	Negative	394	228	123	283
Unarm	Negative	394	228	123	283
Unauthentic	Both	394	228	123	283
Uncivil	Negative	394	228	123	283
Unclean	Both	394	228	123	283
Uncouth	Both	394	228	123	283
Undependability	Both	394	228	123	283
Undependable	Both	394	228	123	283
Underworld	Negative	394	228	123	283
Undignified	Both	394	228	123	283
Uneconomical	Both	394	228	123	283
Unfaithful	Both	394	228	123	283
Unfeeling	Both	394	228	123	283
Unfortunate#3	Negative	394	228	123	283

Word Sense	Category	Overall	Market Level	Firm Level	Industry Level
Ungrateful	Both	394	228	123	283
Unguarded	Negative	394	228	123	283
Unkind	Both	394	228	123	283
Unreliability	Both	394	228	123	283
Unruly	Negative	394	228	123	283
Unsound	Both	394	228	123	283
Unsoundness	Vice	394	228	123	283
Unsureness	Vice	394	228	123	283
Untrained	Negative	394	228	123	283
Untrustworthy	Both	394	228	123	283
Untruth	Negative	394	228	123	283
Uproot	Negative	394	228	123	283
Utterance	Negative	394	228	123	283
Vagabond	Negative	394	228	123	283
Vagrant	Negative	394	228	123	283
Vagueness	Negative	394	228	123	283
Venomous	Negative	394	228	123	283
Vexation	Negative	394	228	123	283
Vile	Both	394	228	123	283
Viper	Negative	394	228	123	283
Vulgar	Both	394	228	123	283
Wait#4	Negative	394	228	123	283
Wanton	Negative	394	228	123	283
Warlike	Negative	394	228	123	283
Warp	Negative	394	228	123	283
Wasteful	Both	394	228	123	283
Wastefulness	Both	394	228	123	283
Wayward	Negative	394	228	123	283
Wee	Negative	394	228	123	283
Weep	Negative	394	228	123	283
Wench	Negative	394	228	123	283
Whine	Negative	394	228	123	283
Whip#2	Both	394	228	123	283
Wickedness	Negative	394	228	123	283
Wily	Negative	394	228	123	283
Wince	Negative	394	228	123	283
Witchcraft	Negative	394	228	123	283
Worrier	Negative	394	228	123	283
Worry#5	Negative	394	228	123	283
Wretch	Negative	394	228	123	283
Wretchedness	Both	394	228	123	283
Writhe	Negative	394	228	123	283