Aalto University
School of Business

Is media just noise? The link between media factors and stock performance

Finance

Master's thesis

Iivari Lappalainen

Pyry Takala

2013

Department of Finance
Aalto University
School of Business

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Authors Iivari Lappalainen, Pyry Takala

Title of thesis Is media just noise? The link between media factors and stock performance

Degree Master of Science (M.Sc.)

Degree programme Finance

Thesis advisor Vesa Puttonen

Year of approval 2013 **Number of pages** 232 **Language** English

Abstract

PURPOSE OF THE STUDY

Interest towards media analytics has increased significantly by both practitioners and academia alike. The hot topic is whether or not qualitative texts contain information relevant to stock financials, and if they do, whether the impact can be used to earn abnormal returns. In order to answer this, we study the impact media factors have on financial metrics in a novel specification that combines all the major media factors in a holistic media model. To transform qualitative texts information into a “sentiment score”, we develop a new methodology to estimate sentiment more accurately than currently prevailing methods.

DATA AND METHODOLOGY

Our study focuses on the S&P 100 constituents between the time period of 2006 and 2011. As a source of qualitative texts, we use major news publications and earnings announcements retrieved from LexisNexis -database using a web scraper program developed for the purpose of this study. We retrieve the financials data for our study using Thomson Reuters Datastream -database.

In order to estimate investor sentiment, we employ both the customary word count, as well as our novel Linearized Phrase-Structure -methodology. For word count, we test the Harvard Psychological -dictionary and a finance-specific dictionary by Loughran and McDonald (2011). As our data is panel in nature, we analyze the correlations in our error terms in line with Petersen (2009), first without clustering and then clustering by firm and by time. We find time-effect in our error terms, and therefore employ a Fama-Macbeth (1973) methodology with clustering done in quarters. To mitigate a methodological choice driving our results, we run our specifications with a multitude of alternative specifications.

RESULTS

We find that Linearized Phrase-Structure (LPS) outperforms the predominant naïve word count methodology. Also, we find that if employing word counts, researchers should employ context dependent dictionaries, such as Loughran and McDonald's (2011). In terms of our main variables, we find that the existing media factors are not mutually exclusive, and impact financial metrics in chorus. Alas, we do not find statistically significant relationship between sentiment and abnormal returns. However, we find a relationship between aggregate market news volume and abnormal returns, and also between sentiment and abnormal volatility. We infer that our findings support limited attention –theory, and provide evidence against market efficiency.

Keywords News volume, sentiment, content analysis, natural language programming, limited attention, behavioral finance, efficient market hypothesis

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijät Iivari Lappalainen, Pyry Takala	
Työn nimi Is media just noise? The link between media factors and stock performance	
Tutkinto Kauppatieteiden maisteri	
Koulutusohjelma Rahoitus	
Työn ohjaaja Vesa Puttonen	
Hyväksymisvuosi 2013 **Sivumäärä** 232	**Kieli** Englanti

Tiivistelmä

TUTKIELMAN TAVOITTEET

Kiinnostus media-analytiikkaa kohtaan on kohonnut viime aikoina merkittävästi. Keskustelun keskipisteessä on ollut kvalitatiivisten tekstien suhde rahoitusmarkkinoihin, ja niiden pohjalta nouseva mahdollisuus ansaita poikkeavia tuottoja. Tutkiaksemme edellä mainittua suhdetta, käytämme uudenlaista tutkimusspesifikaatiota, joka yhdistää kaikki aiemmin tutkitut merkittävät mediatekijät yhdeksi holistiseksi mediamalliksi. Muuntaaksemme kvalitatiiviset tekstit ”sentimenttiarvoiksi”, kehitämme uuden menetelmän sentimentin arviointiin.

LÄHDEAINEISTO JA MENETELMÄT

Tutkimuksemme keskittyy S&P 100 indeksin yrityksiin aikavälillä 2006 – 2011. Kvalitatiivisten tekstien lähteenä käytämme suurimpia kansainvälisiä uutislähteitä sekä yritysten tulosjulkaisuja. Haemme kvalitatiiviset tekstit LexisNexis -tietokannasta käyttäen ”web scraper” -ohjelmaa, jonka olemme rakentaneet tätä tutkimusta varten. Talousdatan haemme Thomson Reuters Datastream -tietokannasta.
Käytämme sijoittajasentimentin arviointiin kahta eri tapaa: perinteistä sanojen laskemiseen pohjaavaa menetelmää sekä kehittämäämme uutta ”Linearized Phrase-Structure” -menetelmää. Sanojen laskemisen kanssa käytämme kahta eri sanakirjaa: Harvard Psykologia -sanakirjaa sekä tutkijoiden Loughran ja McDonald (2011) talousalan sanakirjaa. Lähdeaineistomme on paneelimuotoinen. Havaitsemme otannassa korrelaatiota ajassa, jonka johdosta käytämme nk. Fama-Macbeth (1973) menetelmää, ja ryhmitämme havaintomme kvartaaleittain. Muuttujien ja menetelmien osalta käytämme useita eri määritelmiä vähentääksemme mahdollisuutta, että menetelmävalinta selittäisi tuloksiamme.

TULOKSET

Tuloksemme osoittavat, että Linearized Phrase-Structure (LPS) suoriutuu paremmin sentimentin arvioinnista kuin sanojenlaskentamenetelmä. Lisäksi, tuloksemme näyttävät, että tutkijoiden tulisi suosia kontekstisidonnaisia sanakirjoja, kuten Loughran ja McDonald (2011) sanakirjoja. Muuttujien osalta havaitsemme, että mediamuuttujat eivät ole toisensa poissulkevia. Emme löydä tilastollisesti merkittävää suhdetta sentimentin ja poikkeavien tuottojen välillä. Löydämme kuitenkin tilastollisesti merkittävän suhteen markkinoiden uutismäärän ja poikkeavien tuottojen välillä, sekä sentimentin ja volatiliteetin välillä Tuloksemme tukevat nk. rajallisen huomion (”limited attention”) teoriaa, ja viittaavat siihen, että markkinat eivät allokoi varoja tehokkaasti.

Avainsanat Uutismäärä, sentimentti, sisältöanalyysi, rajallinen huomio, luonnollisen kielen käsittelymenetelmät, rahoituksen käyttäytymistiede, tehokkaat markkinat -hypoteesi

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Background

> "A new breed of forecasters is using '*sentiment analysis'* to pick out the emotionally charged words and phrases which pepper online exchanges."
>
> \- The Economist[1]

Media analytics has experienced a surge in interest from both practitioners and academia alike. Recent academic papers, as well as the wallets of Wall Street investors, have voted for different media factors having an impact on financial metrics. The jury is still out, but the buzz is growing by the minute.

The thought that media factors are linked with financial metrics is based on the idea that qualitative texts have an impact on financial metrics beyond quantitative information. The idea has a two-fold rationale. First, if qualitative information has information content beyond quantitative information, media text can have an impact on financial metrics. Second, if style and tone of text affect the beliefs, preferences and decisions of investors, then qualitative texts can have an impact on financial metrics even in the absence of new information content. The two reasons are not mutually exclusive, and can in fact affect investors' decisions in tandem.

The thought that qualitative texts have information content is based on several findings. First, qualitative texts are thought to be used by managers to communicate adverse information to the public. The argument is that qualitative information is harder to process and can therefore cloud the full impact of adverse information from investors (e.g., Bloomfield, 2002; Davies et al., 2008; Engelberg, 2008). The idea gained wide recognition in the 1990's with the seminal article of Skinner (1994). More recently, for instance, Li (2008) has put forth similar evidence. Second, qualitative texts are thought to be used by managers to communicate future estimates to investors. The rationale is that managers have more freedom in writing qualitative texts which are loosely regulated vis-à-vis quantitative information which is strictly regulated (e.g., Li, 2006; Davis et al,. 2008). In summary, the two categories of

[1] The Economist: Getting in the Mood. The World In 2012 print edition

findings indicate that qualitative information has informational content over and beyond quantitative information.

The idea that qualitative information impacts agents' beliefs, preferences and decisions is rooted in cognitive psychology, and the recent exciting findings of behavioral finance. The key theory behind the impact is '*framing*' which states that agents are impacted by the framing of a problem (e.g., Tversky and Thaler, 1990; Starmer, 2000). As framing is documented to have an impact on agents' behavior, the linguistic style and tone choices of qualitative texts can impact agents' decisions. Indeed, a multitude of studies have found that recipients of messages are attentive to both content and style (e.g., Chung and Pennebaker, 2007). For example, the way stock market commentators describe price movements influences investors' expectations of future prices even in the absence of any fundamental information (Morris et al., 2005).

If qualitative texts can impact investors, they can explain variations in financial metrics. Recent literature has focused on analyzing three different media factors that have been hypothesized to explain variations in financial metrics. First, extant literature has analyzed the impact of firm specific news volume on financial metrics (e.g., Fang and Peress, 2009). Second, prior literature has studied the effect of aggregate market news volume on financial metrics (Hirsleifer et al., 2009). Third, the impact of sentiment on financial metrics has been explored (e.g., Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011). With all the aforementioned variables, findings have documented an impact between the factors and financial metrics.

Until today, the surge in interest has mainly focused on analyzing sentiment's impact on financial metrics. In the epicenter of sentiment studies stands the methodology used to estimate investor sentiment. So far, extant literature has preferred the use of simplistic methodology for the sake of objectivity, replicability and transparency (e.g., Tetlock et al., 2008). The pivotal studies shaping the field of content analysis in finance have relied on word count methodology in combination with a specific dictionary (e.g., Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011). The dominant dictionary used has been the Harvard Psychology dictionary. However, with the seminal article of Loughran and McDonald (2011), that shows that Harvard Psychology dictionary misclassifies words in financial context, the preferred choice of dictionary has become unclear. Also, critique towards the naïve word count methodology has gained footing (e.g., O'Hare et al, 2009). In summary, academia stands in cross-roads: having to decide on the direction that future research will take in terms of methodology.

If qualitative texts impact financial metrics, the question remains whether or not that impact is reflected in the financial metrics according to the propositions of the efficient market hypothesis. The findings of prior literature show that this is not the case. The extant literature has documented both overreaction and underreaction effects. However, the majority of the findings demonstrate an underreaction effect (e.g., Chan, 2003; Loughran and McDonald, 2011).[2] An underreaction effect occurs when information content dominates over tone (e.g., Tetlock, 2007). Therefore, prior literature's evidence seems to indicate that information content is the dominating effect in the impact qualitative texts have on financial metrics.

If underreaction is indeed the dominant effect, scholars must attempt to understand why that is the case: what are the drivers behind the inefficient market causing an underreaction. There are several competing theories. We will introduce three potential explanations. First, '*limited attention*' -theory has been proposed by some scholars (e.g., Hirsleifer et al., 2009) to explain underreaction. Limited attention states that agents have limited cognitive capabilities, and therefore face constraints in information processing depending on the volume and complexity of information. Second, cognitive biases from psychology can explain underreaction.[3] Third, due to the leniency of qualitative text regulations, agents discount information in qualitative texts, leading to underreaction. As uncertainty relating to the information diminishes with time, the discount factor approaches zero, and financial metrics respond correspondingly (e.g., Mercer, 2004; Davis et al., 2008).[4] The outcomes of the aforementioned theories often bear great resemblance to each other. However, the underlying rationales are different.

To sum up, prior literature has presented evidence that qualitative texts have an impact on investors beyond quantitative information. Furthermore, recent literature has presented findings linking different media factors, in particular media sentiment, with variations in financial metrics. The majority of evidence is pointing towards an underreaction effect. To

[2] Tetlock (2007) and Antweiler and Frank (2006) find an overreaction effect with qualitative texts from Wall Street Journal. We hypothesize that the informational content of qualitative texts depends on the source of the qualitative text. Hence, we argue – in line with Tetlock et al. (2008) – that Wall Street Journal articles recapitulate previous news and have no new information content. However, the tone of the articles still affects investors, and therefore causes overreaction that reverses with time.

[3] For instance, two related biases: anchoring and conservatism, can explain underreaction.

[4] Depending on the past context, agents' aforementioned behavior can in fact be rational. Key consideration impacting the discounting is the source credibility: the track record of the author in publishing accurate information via qualitative texts.

explain the effect, academia has proposed several competing theories; no consensus exists at the time of the writing.

1.2 Motivation and definition of research questions

As we briefly mentioned, content literature in the domain of finance stands in a cross-roads. The methodology used to estimate investor sentiment is the center piece of any study focusing on the link between sentiment and financial metrics. Therefore, the choice of investor sentiment estimation methodology should not be taken lightly. Extant literature has preferred the use of simplistic methods in estimating investor sentiment. However, concerns have started to mount up on the validity of such a naïve methodology. Furthermore, the status quo methodology, word counting, is facing a critical choice within it: the choice of preferred dictionary. Before the pivotal study of Loughran and McDonald (2011), the default choice dictionary for academia was the Harvard Psychology dictionary. However, as Loughran and McDonald showed, Harvard Psychology dictionary misclassifies many words in the financial domain. Loughran and McDonald conclude in their paper that scholars should use their context dependent dictionary in future research. However, they add that their dictionary should be tested in a sample consisting of different qualitative texts than 10-k reports[5] to verify that their results hold outside the domain of 10-k reports.

Taking the next step in estimating investor sentiment stands to benefit both academia and practitioners alike. As erroneous estimation methodology results in a measurement error in investor sentiment estimates, the benefits of improving the methodology are obvious. Moreover, Loughran and McDonald (2011) show that erroneous methodology can result in type I errors, and spurious correlations, that can in fact result in incorrect conclusions based on misleading findings. Therefore, improved methodology can change our understanding of the link between investor sentiment and financial metrics from the point-of-view of academia. On the other hand, improved methodology yields more accurate sentiment estimates – something investors find most valuable. In an era where new methodological discoveries

[5] 10-K reports are a commonly used financial reporting format used in the United States. These reports present a summary of a company's performance and are submitted annually to the Securities and Exchange Commission. Typically, the 10-K contains much more detail than a company’s annual report, e.g. detailing out a company’s history, organizational structure, equity, holdings, etc.

result in sentiment funds being raised from nothing to something in an instance,[6] the significance of a breakthrough in investor sentiment estimation methodology is clear.

However, sentiment is not the only media factor of interest. Firm specific news volume (e.g., Fang and Peress, 2009) and aggregate market news volume (e.g., Hirsleifer et al., 2009) have been linked with financial metrics. Yet, no study to date has researched the impact of all the media factors simultaneously. We argue that the factors are not mutually exclusive. Moreover, by exploring all the media factors simultaneously in a holistic media model, one can potentially extract new insights on the impact each of the factors have on financial metrics, and whether or not all of the factors are indeed separate and significant factors of different financial metrics. Furthermore, the insights provided by a study utilizing a holistic media model can be invaluable to the theory building behind the argued relationships between financial metrics and media factors.

Besides the aforementioned methodological and theoretical considerations, the extant literature has so far utilized a narrow scope in event windows and sources of qualitative texts. Therefore, prior literature's findings are subject to critique concerning data dredging. A study employing a large set of event windows with a wide cross-section of qualitative text sources would be free from such critique, and therefore add valuable evidence to the existing literature. The new evidence would shed more light on the impact that qualitative texts have on financial metrics, and on the state of market efficiency.

To sum up, we are interested in finding answers to the following questions:

Methodological questions

- ❖ How can media sentiment be estimated more accurately, using contemporary computer science research? Can a more sophisticated methodology outperform extant naïve methodology? If so, are the results based on the new methodology different from those of the old methodology?
- ❖ What is the preferred dictionary for the status quo methodology? Can the novel Loughran and McDonald (2011) dictionary outperform the Harvard dictionary in a sample consisting of other qualitative texts than 10-k reports?

[6] In 2011, professor Johan Bollen from Indiana University's School of Informatics and Computing published a study linking twitter sentiment with Dow Jones Industrial Average returns. Within days, Johan Bollen had licensed his algorithm to Derwent Capital Management, a hedge fund that incepted a 40 USDm fund based on that sentiment estimation algorithm.

Theoretical questions

- What are the significant media factors driving variations in financial metrics? Is market disseminating information efficiently?
- Can we draw new inferences on the impact that qualitative texts have on financial metrics, and on the theory underlying the impact, based on the analysis of all the media factors simultaneously?

Practical questions

While we do no aim to answer the following in this paper, a paper by Mitra and Mitra (2010) highlights also the following practical questions:

- Trading: could sentiment analysis be used by traders to give insight into questions on which assets to buy, hold or sell?
- Risk control: could information in media and news contain early warning signs that could be important in predicting risk? How could risk managers use media analysis to better understand how different events might impact their portfolio risk?

1.3 Contribution to existing literature

Based on the research questions of previous sub-section, we aim to contribute to the prior literature by:

- Developing a more sophisticated methodology for estimating investor sentiment
- Testing the three documented major media variables on financial metrics simultaneously in an attempt to create a holistic media model, and to isolate the significant media factors that drive variations in financial metrics
- Testing the prior literature's principal methodology for estimating investor sentiment: the vector word count, with the two most prevalent dictionaries utilized: the Loughran and McDonald (2011) dictionary; and the Harvard Psychology dictionary, in order to clarify the efficacy of the extant methodology, and the preferred dictionary for the methodology
- Testing Loughran and McDonald's (2011) dictionary in an out-of-sample test with qualitative texts other than 10-k reports.

- Analyzing the impact different media factors have on financial metrics in a robust study employing a comprehensive cross-section of different qualitative texts with multiple event windows in order to attenuate concerns over data dredging
- Draw conclusions on the theory underlying the link between financial metrics and media factors based on the findings of the study
- Put forward evidence concerning the efficiency of the market, and the informational content of qualitative texts, based on the findings

1.4 Limitations of the study

The limitations relating to our study can be divided into two different broad categories: first, the limitations relating to our sentiment estimation methodology; second, the limitations relating to the data we are using.

Despite the fact that the methodology we have developed for the study: Linearized Phrase-Structure (LPS)[7], is a significant improvement from the naïve word count methodology used in the extant literature, there are caveats that the reader should bear in mind. The main limitations relating to our sentiment estimation methodology are:

- Inability to distinguish topic of a text and its relevance (i.e., the article discusses the target company briefly, or is entirely focused on the target company)
- Failure to assess the credibility of a text source (i.e., local newspaper vis-à-vis Financial Times)
- Inability to differentiate between information concerning: past, present, or future, in a text (i.e., discussion on the historical profit development vis-à-vis speculation on the future profit development of a company)
- Use of only negativity sentiment – both in sentence and article level – to measure sentiment (an alternative would be to use a number of different metrics: for instance, disagreement -sentiment of Das, 2010, might yield additional insights to the link between sentiment and financials)

To overcome the aforementioned limitations would require substantial work, and is outside the scope of this study. We therefore urge future research to focus on the limitations we have mentioned in order to improve sentiment estimation methodology.

[7] Linearized Phrase-Structure -model is our approach for predicting semantic orientations of short economic texts. We define LPS in detail in section 5.

The second category of limitations in our study is data related. The following data limitations should be considered when interpreting our results:

- LexisNexis database coverage of qualitative texts (i.e., not all important qualitative text publications are covered by LexisNexis due to, for instance, copyright issues)
- Missing qualitative texts relating to key products that do not mention the company name (i.e., influential technology paper reviews the new Nokia Lumia phone but does not state Nokia’s name in the article)
- Excluded qualitative sources (e.g., Social media, non-written qualitative media such as television or radio)
- Time period: financial crisis (i.e., unusual patterns in data can exist due to the exceptional circumstances relating to the macroeconomic environment). As discussed later in context of our univariate results, we notice that negative returns combined with higher news volume often concentrated to years 2009-2010
- Daily data instead of intraday data (i.e., we are missing the potential intraday effects from our study)
- Sample firms: S&P 100 (i.e., we are missing small-firm related effects due to our sample consisting of solely large firms). It is possible that the effects we are interested in could be more prominent for small cap companies. Furthermore, the size and news coverage within the S&P 100 varies significantly[8]. Further studies could overcome this limitation possibly by either standardizing news volume per company, or selecting a sample of companies with even more similar news volumes.

Even though our data has limitations, we do not hypothesize that the limitations would significantly affect the nature of our results, or the interpretations we have drawn from our findings.

1.5 Main findings

The findings of our study can be divided into four different broad categories. First, we provide evidence to the on-going methodological debate concerning sentiment estimation. Second, we offer insights on the effects that the three different dominant media factors have on financial metrics. Third, we present findings that contribute to the theory building behind the hypothesized relationship between media factors and financial metrics, and to the

[8] See univariate results in section 6.2.

discussion concerning the information content of qualitative texts. Fourth, we provide evidence on the state of market efficiency.

To test the different methodologies used to estimate investor sentiment, we run a comprehensive benchmarking test. We test the extant literature's prevalent methodology: word count, with Harvard Psychology dictionary and Loughran and McDonald (2011) dictionary, against our novel more sophisticated methodology: the Linearized Phrase-Structure (LPS) -model. The result is clear: Linearized Phrase-Structure -model outperforms the prevalent methodology. Our test utilizes a comprehensive cross-section of qualitative texts according to multiple standard criteria for assessing a classification algorithm's performance. Therefore, we argue that our results are a clear indication that Linearized Phrase-Structure -model improves sentiment estimation significantly. Also, we find that the context specific dictionary of Loughran and McDonald (2011) outperforms the Harvard Psychology dictionary in financial context. As a result, we infer that our benchmarking test offers support to Loughran and McDonald's (2011) claim that academia should prefer their context dependent dictionaries over Harvard Psychology dictionary when assessing sentiment within the financial domain. After establishing the accuracy of used methodology in sentiment estimation, we move to analyze the different media factors prior literature has linked with financial metrics.

We find that investor sentiment is not linked with abnormal returns or volume, but does have a relationship with abnormal volatility. We hypothesize that the noise in prior literature's sentiment estimation methodology has resulted in spurious correlations illustrated by the findings of extant literature in the case of abnormal returns and volume.[9] Also, we find qualitative support that underreaction is the effect related to sentiment and financial metrics.

Next, we confirm the findings of Hirsleifer et al. (2009) by documenting that aggregate market news volume impacts abnormal returns and volume. The impact is characterized by an underreaction effect. We infer that the findings provide strong support for Hirsleifer et al. (2009) proposed *'distraction hypothesis'* that is based on limited attention theory.

Finally, we find that firm specific news volume is not related to abnormal returns. However, firm specific news volume is related to abnormal volume and volatility (only the short event windows are significant with abnormal volume). We infer that noise traders are attracted to

[9] Loughran and McDonald (2011) express similar concerns over spurious correlation when studies employ erroneous sentiment estimation methodologies.

attention grabbing stocks; therefore, increasing the volume and volatility of such stocks. Also, the fact that only the short event windows are significant in terms of volume, offers some support to limited attention theory: as more firm specific news translate to more attention towards a given firm, the lag in information dissemination should decrease. Hence, trades occur more likely close to the event date, and abnormal volumes should only be exhibited in the shorter event windows.

Our findings indicate that the dominant emerging pattern between financial metrics and media factors is underreaction. The implications drawn from the finding are two-fold. First, as underreaction is linked with new information in qualitative texts (e.g., Tetlock, 2007), the evidence supports the assertion that qualitative texts have novel information content. Second, as the relationship between aggregate market news volume is theorized to be related to limited attention, and we are unaware of competing explanations, we infer that limited attention indeed does impact the markets as is indicated by the statistically significant relationship that aggregate market news volume has with financial metrics. Moreover, as other media factors also portray a pattern of underreaction, we suggest that limited attention is the underlying theory explaining the pattern for all the media factors. Our rational is as follows: as limited attention is a phenomenon present in the markets, as demonstrated by aggregate market news volume, its effect cannot exist in isolation from other media factors. Therefore, it should also affect other media factors, and hence have an impact on the relationship other media factors have with financial metrics. We therefore suggest that limited attention is in fact the key underlying theory explaining the underreaction pattern associated with media factors and financial metrics.

In terms of market efficiency, our findings provide valuable evidence to the discussion concerning the state of market efficiency. We find that aggregate market news volume and momentum factors explain variations in abnormal returns in contradiction to the theoretical definition of the efficient market hypothesis. However, we can only infer that the findings are in contradiction to the theoretical definition of market efficiency as we have not tested the economic significance of our results. In other words, we do not know if the relationship is significant ex-post trading costs, and therefore we cannot draw inference on whether or not the findings are in contradiction to the economic definition of market efficiency (e.g., Jensen, 1978) that states that markets are efficient as long as abnormal profits do not persist ex-post trading costs.

Based on our findings, we conclude that future studies could keep on developing more sophisticated methods for estimating sentiment more accurately, especially taking into consideration what qualitative information is most relevant for stock performance. In terms of market efficiency, we suggest that future research should study the effectiveness of specific trading strategies to provide insights to the economic state of market efficiency. Such studies should focus on trading strategies that take advantage of days with high aggregate market news volume. Also, sentiment estimates could be valuable when designing trading strategies. To elaborate more, future research could study a trading strategy that uses sentiment estimates to forecast future stock volatility. Based on the volatility forecasts, the trading strategy could in theory take positions in derivatives to earn alpha. Also, a trading strategy utilizing an index such as the VIX volatility index could prove useful.

1.6 Structure of the study

The rest of the paper is organized as follows. Section 2 reviews prior literature and builds the foundation we will construct our hypotheses on. Section 3 highlights our contribution to the existing literature, and introduces our hypotheses for the study. Section 4 explains the data we are using, the process of gathering it, and presents brief variable descriptions. Section 5 introduces the methodology in the study: the sentiment estimation methodology, and the statistical methods and specifications used to analyze the relationship between our main variables and financial metrics. Section 6 presents the findings of the study, and discusses their impact. Section 7 concludes, and introduces areas we suggest for future research.

2 PREVIOUS LITERATURE REVIEW

The impact of different media factors on financial metrics has spurred up considerable interest during the last decade. Three different major media factors have been linked with variations in financial metrics. In order to understand the implications and context of media analytic studies in the domain of finance, the reader must be acquaint with a vast cross-section of different research topics in finance, psychology and computer sciences. Therefore, we aim to introduce the reader briefly to three major areas that are necessary for the interpretation of our findings. First, we will discuss the efficient market hypothesis that sets the ground for any study in finance researching the link between a variable and abnormal returns. Second, we will introduce the reader to behavioral finance, and the cognitive biases that underlie the nascent field[10]. Third, we acquaint the reader with content analysis research both in finance and other disciplines.

The first part of the literature review will go through the efficient market hypothesis; the modern finance paradigm, and the competing school of thought; behavioral finance, to set the ground for our research. With the birth of behavioral finance, the last few decades have witnessed a fierce debate concerning the state of market efficiency: a debate that, at times, seems to have resembled more a religious quarrel than an academic debate. To illustrate, we offer the following quote from Haugen's[11] (1999) book (Chapter 7, Note 5, p. 71):

> "On April 16, 1998 at the UCLA Conference, The Market Efficiency Debate: A Break from Tradition, while delivering a paper on market efficiency, Fama pointed to me in the audience and called me a criminal. He then said that he believed that *God* knew that the stock market was efficient…"

The quote illustrates an alarming situation: at times, the debate concerning market efficiency seems to have transmogrified into philosophic credence and lies beyond scientific endeavor (e.g., Lee and Yen, 2008). As we hope to provide evidence to the market efficiency discussion, we acquaint the reader with both schools of thought, and the debate between them, to set the context in which our findings will be evaluated.

[10] Behavioral finance is a crucial part of our study, as behavioral finance justifies the rationale as to why qualitative texts can impact agents' behavior even in the absence of new information in qualitative texts.

[11] Haugen is a well-known critic of efficient markets and an advocate of behavioral finance

After reviewing efficient markets hypothesis and behavioral finance theories we will move on to examine the research on content analysis in the domain of finance. We will show that the previous research has demonstrated that by turning qualitative information into quantitative information investors can reap abnormal profits in violation of the efficient market hypothesis.[12] We will focus on reviewing the main findings of previous literature and on demonstrating the different suggested links between theory and investor sentiment.

The section will proceed as follows. First, we will discuss the modern finance paradigm and the efficient market hypothesis. Second, we will briefly go through behavioral finance. Finally, we will explore the previous research in the field of content analysis with a focus in finance.

2.1 Efficient market hypothesis

The literature concerning the efficient market hypothesis is extensive, and a comprehensive review of the literature is a daunting task.[13] Hence, our intention is to introduce the reader briefly to the theory of the efficient market hypothesis [EMH], the history of the hypothesis, and the debate surrounding the EMH, to establish the context for our study and the ensuing discussion on the impact of our results on modern finance paradigm. If the reader is familiar with the EMH literature, he or she may choose to skip the sub-section in question.

The sub-section is organized as follows. First, we will introduce the reader to the EMH, and the different efficiency forms. Second, we will go through the different arguments concerning the debate surrounding the state of the EMH. Finally, we will discuss limits to arbitrage – a topic crucial to both efficient market hypothesis proponents as well as to behavioral finance advocates.

2.1.1 Different forms of efficiency

At the age of 25, Eugene Fama (1965) wrote the following in his seminal PhD thesis:

> "Independence of successive price changes is consistent with an "efficient" market, that is, a market where prices at every point in time represent best estimates of intrinsic values. This implies in turn that, when an intrinsic value changes, the actual price will adjust "instantaneously," where instantaneously means, among other things, that the

[12] However, the violation depends on the definition of efficiency. More discussion on the topic will follow.

[13] For reviews on the EMH, see, for instance: Fama (1991), Lo (1997), Dimson and Mussavian (1998), Farmer and Lo (1999), Beechey et al. (2000), Lee and Yen (2008), and Sewell (2011).

> actual price will initially overshoot the new intrinsic value as often as it will undershoot it."

The concept of an efficient market had been defined[14].

Paul Samuelson provided the first formal economic argument for the EMH (Samuelson, 1965) while Harry Roberts made the distinction between weak and strong form tests (Roberts, 1967) which became the classic taxonomy used with the EMH. Building on the growing evidence regarding the EMH, Fama published his seminal review of the efficient market theory and evidence (Fama, 1970). In his article, Fama (1970) defined the efficient market to be categorized into three types of efficiency: the weak form, the semi-strong form and the strong form. However, Fama (1991) later changed the categories. The weak form tests was renamed to tests for return predictability, the semi-strong form tests was relabeled to event studies, and the strong form tests was renamed to tests for private information. In spite of the change to categories, we will use the original taxonomy, and the characterization relating to it, as it is more widely recognized and used.

Fama (1970) defined the market to be in line with the weak form efficiency if the prevailing prices reflect all historical prices. At that time, a competing group of theories, the chartist theories,[15] suggested that past prices contained significant amounts of information concerning future prices. However, the empirical evidence was in line with the weak form efficiency and the chartist theories were quickly abandoned from the academic research. In 1991, Fama changed the name of the weak form tests into tests for return predictability and included in the category such variables as dividend yield and interest rates in addition to historical prices.

Under the semi-strong form efficiency, prices will fully reflect all available public information and adjust to any new information instantaneously and in an unbiased manner (Fama, 1970). In other words, overreactions will be as common as underreactions if they occur. Later on, Fama (1991) changed the category name to event studies in the wake of the large event study research spurred up by the first event study in 1969 by Fama, Fisher, Jensen and Roll.[16]

[14] For history of the efficient market hypothesis, we refer the reader to section 0.

[15] The most well-known of the chartist theories is probably the Dow Theory which is often cited as the origin of technical analysis.

[16] In fact, the first published event study was by Ball and Brown (1968), however, Fama et al. (1969) was the first event study undertaken.

Strong form efficiency is defined to fully reflect all available information so that no individual has higher expected trading profits than others because of monopolistic access to some information (Fama, 1970). In other words, all private information is fully reflected in the prevailing prices. However, Fama (1970) states that the strong form efficiency is in reality an extreme null-hypothesis; a clean benchmark, for tests against market efficiency. In fact, such groups as corporate insiders do have monopolistic information and hence prove that the strong form efficiency is not a strictly valid theory. Consequently, the strong form efficiency is not expected to be an exact description of reality but instead can be used to find out how far down through the investment community do the deviations permeate from the strong form efficiency. In 1991, Fama relabeled the strong form efficiency into tests for private information. However, the content of the category remained the same.

2.1.2 Debate concerning the state of the efficient market hypothesis

In the zenith of its time, during the 1960s, most of the evidence was in favor of the EMH. With few exceptions such as Cowles (1960), Niederhoffer and Osborne (1966) and Scholes (1969), all evidence seemed to support the EMH.[17] However, as time passed, evidence contradicting the EMH began to mount up and concerns were raised that the EMH was flawed.

In 1978, a special edition of the Journal of Financial Economics (Vol. 6, Numbers 2 to 3, June/September) was devoted exclusively to the anomalies surrounding the EMH. The special edition was the overture for the forthcoming variety of anomalies that would cast doubt on the EMH. Such anomalies included:[18]

- ❖ *Small firm effect:* Small-capitalization firms earn higher than average returns vis-à-vis their expected return based on different asset pricing models (e.g., Banz, 1981; Reinganum, 1981; Keim, 1983; Brown et al., 1983; Schwert, 1983; Fama and French, 1993; Rouwenhorst, 1999)

[17] Lee and Yen (2008) hypothesize that the seemingly scarce evidence against the EMH during the 1960s might not be as scarce as it might first appear due to four following reasons. First, the scientific paradigm proposed by Kuhn (1970) might act as a protective belt for the EMH. Second, testing bias was present as reported by LeRoy (1989). Third, improper statistical methods were used as suggested by Taylor (1982). Fourth, the empirical evidence might have been misinterpreted to support EMH when in fact it was against it, as pointed out by Arbit and Boldt (1984), and Lee and Yen (2008), in the case of Fama et al. (1969).

[18] The following list is incomplete and excludes several well-known anomalies (e.g., stock splits). However, the aim is to simply illustrate the voluminous number of anomalies related to the EMH.

- *Value (price-earnings ratio) effect:* Firms with low price-earnings ratios outperform their counterparts with high price-earnings ratios (e.g., Nicholson, 1960; Basu, 1977; Ball, 1978; Basu, 1983; Campbell and Shiller, 1998; Rouwenhorst, 1999)

- *Dividend effect:* Dividend yields, initiations and omissions forecast future abnormal returns (e.g., Long, 1978; Charest, 1978; Rozeff, 1984; Shiller, 1984; Fama and French, 1988b; Campbell and Shiller, 1988; Michaely et al., 1995)

- *Seasoned equity offering effect:* Companies issuing seasoned equity underperform their peers in the long-run (e.g., Loughran and Ritter, 1995; Spiess and Affleck-Graves, 1995; Teoh et al., 1998; Brav et al., 2000; Jegadeesh, 2000; Ritter, 2003)

- *Share repurchases effect:* Positive long-term abnormal profits persist when firms tender for their shares or initiate an open market repurchase program (e.g., Lakonishok and Vermaelen, 1990; Ikenberry et al., 1995; Grullon and Michaely, 1998; Ikenberry et al., 2000)

- *Earnings announcements:* Stock prices tend to respond to earnings with a substantial delay (e.g., Ball and Brown, 1968; Foster et al., 1984; Rendleman et al., 1987; Freeman and Tse, 1989; Bernard and Thomas, 1989 and 1990; Brown and Pope, 1996)

- *Closed-end fund discount effect:* Closed-end funds commonly trade in organized secondary markets at a discount relative to their net asset value (e.g., Thomson, 1978; Lee et al., 1991; Pontiff, 1996; Shleifer, 2000)

- *Excess volatility:* Stock prices experience greater volatility that can be explained by fundamentals (e.g., Shiller, 1981; LeRoy and Porter, 1981; Shiller, 1982; Grossman and Shiller, 1981; Campbell and Shiller, 1988; West, 1988; Shiller, 1992; Hansen and Jagannathan, 1991)

- *Day-of-the-week effect:* The day of the week can be used to forecast future returns, whether it be a Monday, a day around holidays, or an end-of-month day (e.g., Cross, 1973; French, 1980; Gibbons and Hess, 1981; Lakonishok and Levi, 1982; Keim and Stambaugh, 1984; Rogalski, 1984; Smirlock and Starks, 1985; Ariel, 1987; Lakonishok and Smidt, 1988; Ariel, 1990; Hawawini and Keim, 1995)

- *January effect:* The turn-of-the-year is followed by abnormal positive returns in January (e.g., Rozeff and Kinney, 1976; Keim, 1983; Reingaum, 1983; Haugen and Lakonishok, 1988; Ritter, 1988; Haugen and Jorion, 1996; Haug and Hirschey, 2006)

- *Reversal effect:* When shares are ranked on basis of their past returns during three to five years, past winners tend to become losers and vice versa (e.g., DeBondt and Thaler, 1985; DeBondt and Thaler, 1987; Fama and French, 1988a; Poterba and Summers, 1988; Chopra et al., 1992; Richards, 1997)

- *Momentum effect:* Past year's winners tend to outperform past year's losers during the next three to six months (e.g., Jegadeesh and Titman, 1993; Fama and French, 1996; Campbell et al., 1996; Brennan et al., 1998; Rouwenhorst, 1998; Rouwenhorst, 1999; Lo and MacKinlay, 1999; Lo et al., 2000; Jegadeesh and Titman, 2001; Lewellen, 2002)

The commanding empirical evidence inconsistent with the EMH spawned a new school of thought in finance during the 1990s under the nomenclature of behavioral finance. While several of the reported anomalies have perished under scrutiny from the proponents of the EMH, others have survived the ordeal. As a consequence, the nascent school of Behavioral Finance asserts that the inefficiency of capital markets is the norm, not the exception, and as such, the EMH serves as a description of an ideal world; not the real world. Instead the behavioral school advocates the use of more eclectic approaches (e.g., Shiller, 2003). However, as striking as the evidence against the EMH may appear, the advocates of the EMH have evidence of equal proportions to cast at the zealots of behavioral finance. The debate is fierce and ongoing.

The followers of EMH argue that the behavioral finance school runs head on to the enigmatic joint-hypothesis problem first introduced by Fama (1970). The joint-hypothesis of the efficient markets has puzzled researchers for decades. The implication of the joint-hypothesis is that when testing for market efficiency; how prices reflect available information, one must also test simultaneously for an asset pricing model (i.e. the Sharpe-Lintner-Black model [CAPM] of Sharpe (1964), Lintner (1965) and Black (1972)). Hence, one can test market efficiency conditional to an asset pricing model or asset pricing model conditional to market efficiency. Therefore, inherently, all tests of market efficiency are at the same time tests of an asset pricing model. As a consequence, the supporters of the EMH have argued that the anomalies deviating from the EMH are in fact '*bad-model problems*' (e.g., Fama, 1991, 1998;

Malkiel, 2003) produced by misspecified asset pricing models. Furthermore, with long-term returns, the bad model problem escalates, as the asset pricing model plays a larger role in measuring abnormal returns than it does in the short-term.[19]

Proponents of the EMH have also raised concerns over data mining as a source of spurious regularities (e.g., Fama, 1991, 1998; Schwert, 2003; Malkiel, 2003). Given the current state for availability of data with sophisticated databases like CRSP, the concern is indeed not without merit. Furthermore, as suggested e.g. by Merton (1985), exciting new findings get reported while the outcomes that confirm the norm never see the light of day (See: Fama, 1998; Schwert, 2003 and Malkiel, 2003). As the EMH is the norm, the vast empirical evidence inconsistent with the EMH is in fact far from a random sample and as such suffers from a severe case of sample bias.

As samples can exhibit specific patterns that can lead to spurious regularities when data mining techniques are employed, out-of-sample tests are warranted. The proponents of the EMH argue that such tests refute the majority of anomalies (e.g., Fama, 1998). Furthermore, even if out-of-sample tests would confirm an anomaly, the issue of time dependence; whether or not the anomaly persists over time, is not resolved (Malkiel, 2003). A case in point, as demonstrated by Schwert (2003), is the Dimensional Fund Advisors (DFA) that incepted mutual funds targeted to exploit discovered anomalies (i.e. size- and value-effect). However, the DFA mutual funds have not succeeded in generating positive abnormal returns. Therefore, the anomalies in question seem to have dissipated, or failed the test of economic significance. Schwert (2003) offers a different view: when anomalies are discovered they simultaneously disappear as practitioners employ strategies to take advantage of the anomalies. Hence, research findings cause the market to become efficient. Grossman and Stiglitz (1980) put forward a similar view arguing that markets cannot be completely efficient as otherwise there would be no incentive for professionals to uncover information. Thus, they build on Schwert (2003) by assigning the discovery function to professionals.

In the event that an anomaly persists in out-of-sample tests during different time periods, the conclusion is still ambiguous. Fama (1998) argues that anomalies tend to disappear when reasonable alternative approaches are used. For instance, Fama argues that by switching the asset pricing model more than a few of the anomalies disappear. Also, by altering the return

[19] With short-term studies, the expected return approaches 0 and the asset pricing model plays a smaller role whereas with long-term studies the expected return is affected by the model to a greater extent.

metric from buy-and-hold abnormal returns [BHARs] to average abnormal returns [AARs], or cumulative abnormal returns [CARs], mitigates several of the anomalies. Likewise, by implementing value weights in place of equal weights a number of the anomalies are erased.

In addition, there are countless anomaly specific debates on-going. For instance, concerning the small firm -effect, Malkiel (2003) argues that the well-documented anomaly is in fact a result of survivorship bias as the modern computerized databases include only small firms that have survived; not the ones that went bankrupt.[20] Therefore, the apparent anomaly is a result of a sample bias.

If an apparent anomaly is accepted as a real anomaly, there are still issues that need to be resolved. First, Fama (1998) maintains that anomalies are chance results: apparent underreactions will be about as frequent as overreactions. If so, the market is in fact efficient. Shiller (2003) counters Fama's proposition by stating that there is no fundamental psychological principle that people tend always to overreact or underreact, hence it is no surprise that research on financial anomalies does not reveal such a principle either. For that reason, the random split is by no means proof of market efficiency.[21] Second, the proponents of the EMH reason that if the EMH is to be replaced, it can be done only by a better specific model of price formation (e.g., Nichols, 1993; Fama, 1998; Schwert, 2003). Fama (1998) argues that as the behavioral finance school has not tested consistently a specific alternative asset pricing model, the EMH cannot be rejected as there is no alternative. Furthermore, Fama continues by claiming that the existing behavioral models do not produce rejectable predictions that capture the menu of anomalies, instead they capture the anomaly they are designed to capture failing disastrously with other anomalies. In Schwert's (2003) opinion, the future models must go beyond explaining discovered anomalies to improve our understanding of asset pricing.

In addition to the documented anomalies, recent financial crises have expedited the spiraling decay of the EMH. Many former EMH scholars have jumped ship stating that such prodigious prices cannot be explained by rational valuations under any circumstances. Consequently, the role of the market as an efficient capital allocator has been questioned widely. The supporters of behavioral finance have reasoned that irrational bubbles are in fact the product of feedback

[20] Small firms tend to be more highly levered which could lead to, or be a consequence of, financial distress (e.g., Bhandari, 1988).

[21] We would also like to take the opportunity to remind the reader that, in any circumstances, the absence of proof is not proof of absence.

models[22] (e.g., Smith et al., 1988; Shleifer, 2000; Shiller, 2000a, 2003). In contrast, the efficient markets proponents maintain that the bubbles are in fact large systematic patterns in the variation of expected return through time (e.g., Fama and French, 1989; Fama, 1991; Malkiel, 2003). For instance, Miller (1991) explains the market crash of October 1987 as follows: external events, minor in themselves, could have cumulatively signaled a possible change in what had been up to then a very favorable political and economic climate for equities. Therefore, the crash was in fact a result of a rational shift in equity prices to accommodate the change in expected returns. Nothing is as clear in prospect as it is in retrospect. Nevertheless, irrespective of some of the differing views concerning the reasons for crises, both sides agree on the importance of market frictions in their formation; in other words, the importance of limited arbitrage opportunities (e.g., Shleifer, 2000; Shiller, 2003; Malkiel, 2003; Schwert, 2003).

2.1.3 Limits to arbitrage

The concept of '*limits to arbitrage'* has grown into a major topic in finance. Following the seminal articles of Miller (1977and Jensen (1978):, the surge in interest has been substantial. In fact, in 2002, the Journal of Financial Economics released a special issue dedicated to the limits of arbitrage (Vol. 66, Numbers 2 to 3, November/December 2002). Besides short-selling and transaction costs, limited arbitrage has been associated with noise-trader risk (e.g., De Long et al., 1990; Shleifer and Vishny, 1997; Shleifer, 2000) and fundamental risk (e.g., Barberis and Thaler, 2003). The key argument relating to limited arbitrage is that market frictions, whatever they might be, prevent arbitrageurs from undertaking arbitrage; as a result, prices can deviate from their fundamental values. Therefore, limited arbitrage contradicts Friedman's (1953) seminal assertion that rational traders will quickly undo dislocations caused by irrational traders.

Regarding short-selling constraints, arbitrageurs are unable to short the stock profitably and therefore cannot profit from their knowledge (e.g., Miller, 1977; Jones and Lamont, 2002).[23] As a result, prices can deviate from their fundamental values and anomalies can exist. [24] In the

[22] Feedback models date back all the way to the Dutch Tulip mania.

[23] The cost of shorting has also been associated with Kahneman and Tversky's (1979) prospect theory. The rationale is that psychological factors influence arbitrageurs to avoid shorting due to fear of psychological anguish produced by the need to cover an unprofitable short position (Shiller, 2003).

[24] A case in point is the 3Com sale of Palm Pilot in the early 2000 (e.g., Lamont and Thaler, 2001). 3Com executed a 5% carve-out of its subsidiary's Palm Pilot's outstanding shares. As the shares began to trade, the implied value of the 95% ownership of Palm Pilot by 3Com exceeded the market value of 3Com. As a result, the situation implied a negative value for the rest of the 3Com's business as a whole. The existence of arbitrageurs

case of transaction costs, the argument is simple: anomalous profits can exist statistically but not economically. In other words, the anomalies will lack economic significance after transaction costs are included (e.g., Jensen, 1978; Shleifer, 2000; Barberis and Thaler, 2003).

Concerning noise-trader risk, arbitrageurs are unable to determine the duration of a bubble. As arbitrageurs often face short-term investment horizons, they are unwilling to bear the risk of offsetting noise-traders in fear of that their time-horizon will lapse before the price will revert to fundamentals (e.g., Shleifer, 1990; De Long et al., 1990; Shleifer and Vishny, 1997; Shleifer, 2000; Malkiel, 2003)[25].

On the subject of fundamental risk, arbitrageurs are unable to hedge their position with a similar security due to the fact that substitute securities are rarely perfect. As a consequence, arbitrageurs are left with a proportion of fundamental risk that can effectively limit arbitrage (e.g., Barberis and Thaler, 2003). Wurgler and Zhuravskaya (2002) elegantly demonstrate, in the context of index inclusions, how difficult it actually is to find a good substitute security for an individual stock, therefore exemplifying how important fundamental risk is in arbitrage.

Limits to arbitrage are widely accepted by both EMH and behavioral finance proponents. The discussion has focused mainly on whether or not the limits are against market efficiency. Behavioral finance scholars have maintained that limited arbitrage undercuts the EMH and is a vital part of behavioral finance (e.g., Shleifer and Summer, 1990; Shleifer, 2000; Barberis and Thaler, 2003; Ritter, 2003). However, the supporters of the EMH have argued that limited arbitrage is not against the EMH. They maintain that an economically more accurate hypothesis (e.g., Jensen, 1978) is not in contradiction of the EMH but in fact a realistic representation of the EMH. In reality, the clean benchmark where all information is available and no transaction costs exist is not a truthful illustration of the real world (Fama, 1991). Therefore, anomalies that do not pass the test of economic significance: when marginal profits

should have prevented such a radical deviation, as the market value of any business can be, at worst, zero due to limited liability. However, arbitrageurs were unable to short the Palm Pilot stock due to extremely high borrowing costs. Therefore, the price anomaly persisted until 3Com spun-off more of Palm Pilot's shares, alleviating the mismatch between supply and demand.

[25] An example of the risk is the notorious Royal Dutch and Shell relative arbitrage trade (e.g., Froot and Dabora, as1999). In 1907, Royal Dutch and Shell agreed to merge their interest on a 60-40 basis and pay dividends according to the same basis. Therefore, under modern finance theory, whenever the stock prices are not in 60-40 basis an arbitrage profit opportunity exists. Nevertheless, historically, the basis has deviated significantly from 60-40. In fact, Royal Dutch/Shell arbitrage trade can be viewed as one of the most popular equity arbitrage trades undertaken by the trading desks of Wall Street in the recent history - indeed even the notorious Long-Term Capital Management had the trade in their books at the time of their collapse in the fall of 1998 (Lowenstein, 2002).

of acting on information exceed the marginal costs, are not in fact anomalies at all. The proponents of the EMH have argued that the vast majority of apparent anomalies fall prey to the aforementioned (e.g., Fama, 1970, 1998; Malkiel, 2003). The view of some of the EMH supporters is excellently encapsulated by Richard Roll’s response to Robert Schiller’s comments concerning market inefficiencies at a symposium (Roll and Shiller, 1992)

> “I have personally tried to invest money, my client’s money and my own, in every single anomaly and predictive device that academics have dreamed up… I have attempted to exploit the so-called year-end anomalies and a whole variety of strategies supposedly documented by academic research. And I have yet to make a nickel on any of these supposed market inefficiencies… a true market inefficiency ought to be an exploitable opportunity. If there’s nothing investors can exploit in a systematic way, time in and time out, then it’s very hard to say that information is not being properly incorporated into stock prices.”

To some extent, the term ‘*efficiency*’ has become blurred: on one side, researchers stress that anomalies should be exploitable in the sense that they would yield abnormal profits; on the other side, scholars argue that an anomaly is real if it dictates that prices deviate from fundamentals, hence economy is not allocating assets efficiently to the best investment opportunities (e.g., Statman, 1999; Barberis and Thaler, 2003). As a result, the argument is pinned down to whether or not efficiency is viewed from an ‘*investor*’ or an ‘*economist*’ point of view. Therefore, one might say that opposing camps are viewing the different sides of the same coin. In fact, Brav and Heaton (2002) demonstrate that the behavioral models and rational models describing efficient markets are, in terms of predictability and mathematics, practically indistinguishable from each other even while having very different assumptions concerning the world they model. Whether or not the on-going debate is more about semantics than substance is for the reader to find out.

2.2 Behavioral finance

> “The efficient market hypothesis is the most remarkable error in the history of economic theory.”
>
> - Lawrence Summers, U.S. Secretary of Treasury, on Black Monday crash of 1987

The tide was turning for the modern finance paradigm. The growing amount of empirical evidence in contradiction to the efficient market hypothesis was stirring up debate concerning the validity of the efficient markets hypothesis in the academic society; behavioral finance was born. [26]

In order to understand the debate concerning the validity of the efficient market hypothesis, the reader must understand the arguments of the opposing side: behavioral finance scholars. Also, to understand the argument that qualitative texts can impact agents even in the absence of new information content, the reader must know the underlying cognitive biases that are the basis of the argument. In this section, we introduce the reader briefly to the nascent field of behavioral finance, and the underlying cognitive biases discovered by psychologists.

The roots of behavioral finance can be traced back all the way to John Maynard Keynes’s reference to ‘*animal spirits*’ in 1936. Keynes stressed the role of uncertainty and confidence in shaping the economic life. In Keynes’ view, economic agents’ psychology could be easily manipulated and disturbed, hence underlining the importance of psychology in the economic system. In fact, an argument can be made that behavioral finance is the vindication of Keynesian ideas.[27] However, behavioral finance does diverge from Keynesian tradition in its emphasis of experimental and empirical evidence and the use of formal models to derive predictions. Indeed, behavioral finance can be seen to be close in spirit to Keynesian tradition but with its own methodology and analytical framework (Stracca, 2004).

To better understand the context of behavioral finance, it is appropriate to briefly go through the notions underlying prevailing modern finance paradigm. The modern finance approach posits that agents are rational. By rational, the approach assumes two features: Firstly, agents react to new information correctly updating their beliefs according to Bayes’ law. Secondly, given those correct beliefs, agents act rationally to maximize their expected utility using either

[26] For discussion relating to the anomalies and their interpretation, see the efficient market hypothesis -section.

[27] See Harvey (1998) for a discussion on the Keynesian concepts of uncertainty and non-ergodicity in economic life vis-à-vis modern economic psychology. For a discussion on manias and panics in financial markets relating to Keynesian tradition, see Kindleberger (1978).

objective probability distribution according to the expected utility theory (Von Neumann and Morgenstern, 1944) or subjective probability distribution in line with the subjective expected utility under the notion of Savage (1964).[28] The initial wealth and preferences of agents do not matter as long as financial markets are efficient (Constantinides, 1982). As a result, prices will reflect fundamentals: the phenomenon coined the efficient market hypothesis.[29] The appeal of the modern finance approach is the simplicity and the superior analytical tractability. However, as we have shown in the efficient market hypothesis section (2.1.), the empirical evidence contradicting the prevalent paradigm is vast and covers large ground; aggregate stock markets, cross-sections of average returns as well as individual trading behavior.

Behavioral finance differs from the efficient market hypothesis by arguing that the behavior of agents deviates from the rational behavior dictated by maximization of expected utility (e.g., Starmer, 2000; Barberis and Thaler, 2003; Ritter, 2003; Stracca, 2004).[30] Indeed, a vast literature in the field of psychology has documented numerous cognitive biases that affect the behavior of humans[31] causing them to form beliefs, and act in contradiction to Bayesian law. Hence, agents exhibit, what is characterized as irrational behavior by modern finance, because of mistaken beliefs and preferences. For instance, the impact of emotional and visceral factors in individual decision making is well documented (e.g., Lowenstein, 2000 and Romer, 2000). However, the fact that emotional factors affect decision making is not proof of irrational behavior per se: the question is whether or not emotional responses deviate from rational responses - the answer is unequivocally: yes. Despite the aforementioned, modern finance ignores the impact of emotions on agents' behavior, as well as other cognitive biases and their impact. Modern finance proponents have rebutted the importance of cognitive biases on the basis of learning. The argument is that through repetition agents will learn their way out of biases. In addition, modern finance proponents suggest that strong enough incentives will remove biases altogether. Barberis and Thaler (2003) reason that while learning and incentives can attenuate biases it is unlikely that they will completely remove them. Camerer and Hogarth (1999) point out that while incentives can reduce biases no study has made rationality violations disappear altogether by raising incentives. In fact, the whole concept of

[28] E.g., Barberis and Thaler, 2003; Stracca, 2004.

[29] For mathematical description, we refer the reader to, for instance, Stracca (2004).

[30] To be precise, behavioral finance advocates more eclectic models than those based on Von Neumann - Morgenstern (1944) expected utility (EU) or Savage's (1964) notion of subjective expected utility (SEU).

[31] See e.g., Kahneman et al., 1982; Camerer, 1995; Rabin, 1998; Kahneman and Tversky, 2000; Gilovich et. al., 2002.

learning is dubious if learning implies a painful loss of self-esteem and the recognition of intellectual inferiority in comparison to peers; issues demonstrated to have influence over agents' behavior in experimental psychology (Griffin and Tversky, 1992). To summarize, the argument of mainstream finance that learning will eliminate individual agents' cognitive biases is unresolved.

Nonetheless, agents' irrational behavior is by no means proof of irrational markets. Even in the existence of irrational behavior, the aggregate market would be rational. For irrational traders, '*noise traders*', to be driven out of the market, rational traders, '*arbitrageurs*', must be able to correct prices (Friedman, 1953). However, as we have demonstrated, limited arbitrage prevents this and arbitrageurs are often unable to exploit opportunities created by noise traders.[32] Hence, prices can in fact deviate from their fundamental values.

In order to say more about the structure of the deviations, behavioral finance turns to the extensive experimental evidence on cognitive biases compiled by psychologists. However, mainstream finance theorists argue that the extensive evidence is a double edged sword: a Pyrrhic victory[33] for behavioral finance at best. The predicament is that the vast amount of cognitive biases provide behavioral scholars with so many degrees of freedom that they can explain practically anything with their models - a phenomenon often dubbed '*model dredging*' (e.g., Ritter, 2003). In other words, there is always a story that fits the evidence ex-post.[34] However, opposing views have been presented: Barberis and Thaler (2003) maintain that mainstream theorists have an equivalent amount of flexibility. As Arrow (1986) vehemently argued, rationality per se does not yield many predictions. In fact, the predictions in rational models are often based on auxiliary assumptions, hence offering additional degrees of freedom for modern finance theorists. Barberis and Thaler (2003) conclude by urging both sides to test their theories empirically: especially the assumptions of the theories.[35]

[32] We refer the reader to 2.1.4: Limits to Arbitrage, under 2.1 Efficient Market Hypothesis -section

[33] A victory with devastating cost to the victor; it carries the implication that another such victory will ultimately cause defeat. The origins of the saying originate from the ancient Pyrrhic war fought by King Pyrrhus of Epirus.

[34] See, for instance, Hirshleifer (2001) on the topic of making ex-ante predictions on which model dominates

[35] The emphasis of testing assumptions is directed more towards modern finance theorists: since the influential argument of Milton Friedman, to evaluate theories based on the validity of their predictions rather than their assumptions, testing assumptions has been neglected.

The remainder of the sub-section is organized in two parts: First, we will discuss the cognitive biases unearthed by psychologists. Second, we will discuss the most prominent non-expected utility theory: the prospect theory and its successions.[36]

2.2.1 Cognitive biases

Cognitive capabilities of humans are limited. In an effort to guarantee survival throughout the human evolution, the brain has adapted to solve complex problems in a manner that optimizes deliberation cost and outcome (e.g., Arruñada, 2008). However, as the modern environment has transformed rapidly around us, the brain has become maladapted to the current surroundings. Hence, humans: '*agents'*, can demonstrate seemingly irrational behavior at times due to the misaligned optimization of deliberation vis-à-vis outcome. For instance, agents employ heuristics, rules of thumb, based on past experience to solve a problem.[37] Indeed, heuristics can still be a superior approach to solving problems (e.g., Langlois, 2003). However, that is not always the case. Heuristics simplify a multifarious world, consequently resulting in biased decisions when an agent is confronted with a complex problem (e.g., Stracca, 2004). By a complex problem, we mean information ambiguous to the agent, such as statistical information (e.g., Arruñada, 2008). In fact, even individuals trained in statistics use heuristics when making decisions requiring statistical reasoning (Tversky, 2004), leading to irrational behavior. A case in point is the 1/N allocation rule of retirement funds to different asset classes (e.g., Benartzi and Thaler, 2001). Agents seem to allocate 1/N portion of their retirement funds to the available N asset classes without respect to the actual underlying assets they are investing. For instance, when asked to allocate assets between equity and debt funds, agents will allocate assets on a 50-50 basis. However, when asked to allocate between equity fund and balanced debt-equity fund[38], agents will again allocate assets based on a 50-50 ratio. Therefore, allocating 75% to equity in the latter example vis-à-vis the 50% allocated in the former example: an obvious example of irrational behavior.

To touch upon the most common cognitive biases in more detail, we will briefly describe the ones most commonly connected to behavioral finance. To our knowledge, no universally accepted categorization of the biases exists. Hence, the following categorization is our own and therefore subjective.

[36] For a review on the most dominant behavioral asset pricing models, we refer the interested reader to the works of Barberis et al. (1998), Daniel et al. (1998,2001), Hong and Stein (1999) and Barberis and Shleifer (2003).

[37] See, for instance, Simon, 1986; Williamson, 1997; Kahneman and Tversky, 2003; Arruñada, 2008.

[38] The mixed fund is assumed to hold equity and debt in 50-50 basis.

Overconfidence, self-attribution, hindsight and optimism

Overconfidence is one of the most documented cognitive biases. According to Daniel et al. (1998), overconfidence is defined as overweighing ones private information signals but not the public information signals. Hence, agents are overconfident about their abilities and rely excessively on personal judgment. For instance, overconfidence manifests itself as too narrow confidence intervals in estimates (e.g., Alpert and Raiffa, 1982) and ill-equipped probability estimates; events estimated certain only occur circa 80% of the time, and events estimated improbable actually occur approximately 20% of the time (e.g., Fischhoff et al., 1977). In fact, Odean (1998a) argues that overconfidence leads agents to interpret information in a distorted manner: overweighing salient and anecdotal information while ignoring abstract and statistical information[39]. Kahneman (2003) expands the argument stating that agents become overconfident especially in low-information environment and Tversky (2003) demonstrates that overconfidence takes place regardless of the level of expertise in the subject matter.[40] Also, Barber and Odean (2001) illustrate that men are more overconfident than women exhibited by excessive trading. Indeed, overconfidence is thought to explain excessive trading behavior as agents believe to have superior information and intellectual capabilities vis-à-vis their peers[41]. In fact, overconfidence has been argued to explain several phenomena in the field of finance. For instance, investors diversify their portfolios much less than would be recommended by normative models, and exhibit a strong '*home bias*' in their investments.[42] In the field of corporate finance, Roll (1986) suggests that takeover activity evidence displays a pattern of overconfidence and optimism in managers' decision making: managers have too rosy views on suggested assumptions for the synergy calculations materializing. Hence, takeover bids are often too high vis-à-vis the implied fundamental values and too many takeovers get done in comparison to what should occur in rational markets. Roll (1986) christens the phenomenon '*the hubris hypothesis*'.

Overconfidence is often linked closely to self-attribution and hindsight biases (Barberis and Thaler, 2003). Self-attribution causes agents to credit themselves for the past successes while failures are attributed to external factors such as bad luck. Hirshleifer (2001) demonstrates that when agents receive positive feedback on their private information, agents' confidence

[39] A phenomenon common to several biases

[40] Daniel et al. (1998) show that several professions requiring expertise succumb to overconfidence. Such professions include: lawyers, investment bankers, engineers, psychologists etc.

[41] See, for instance, DeBondt and Thaler, 1995; Odean, 1998b, 2000; Barber and Odean; 2000, 2002a

[42] See, for instance, French and Poterba, 1991; Lewis, 1999; Grinblatt and Keloharju, 2001; Huberman, 2001

increases excessively whereas receiving negative feedback will not sufficiently decrease agents' level of confidence. In other words, agents overweight positive feedback on their private information while underweighting negative feedback, resulting in unrealistic belief of superiority and talent: overconfidence.[43] Another bias that is intertwined with overconfidence is the hindsight bias. Hindsight bias causes agents to believe that they have forecasted an event ex-post the event, when in fact they might have done no such thing. Therefore, hindsight bias results in overconfidence in one's ability to forecast future (e.g., Fisher and Statman, 2000).

Optimism is another bias closely related to the aforementioned biases. Whereas self-attribution and hindsight biases can be seen to be directly linked with overconfidence, optimism is often documented to influence an agent in chorus with overconfidence; for instance, the hubris hypothesis (Roll, 1986) links overconfidence with optimism. Optimism causes agents to display unrealistically rosy views of their abilities and prospects - for instance, over 90% of those surveyed think they are above average in such domains as driving skill (Weinstein, 1980). Buehler et al. (1994) document that agents illustrate a systematic planning fallacy: they predict shorter completion times for tasks vis-à-vis the actual completion times.[44] In the field of corporate finance, optimism has been coined with the famous pecking order theory. As managers hold rosy views on the prospects of their company, they consider equity to be undervalued by the markets. Overconfident on their beliefs, they avoid issuing equity at all cost even to the point of dismissing lucrative investments in order to avoid issuing equity (e.g., Malmendier and Tate, 2005; Heaton, 2002).

Representativeness: law of small numbers & conservatism, and anchoring

Representativeness refers to heuristics that influence agents' behavior. Two opposite heuristic rules apply when agents determine the representativeness of a set of data. The first rule is the so called '*law of small numbers*' which theorizes that agents tend to overweight recent events; '*the sample rate*', at the expense of prior events; '*the base rate'* (e.g., Gilovich et al., 1985; Rabin, 2002). In other words, agents tend to infer from too few data points the actual population parameters. Gilovich et al. (1985) present an illustrative real life example: the so called '*hot hand*' phenomenon. In basketball it is commonly believed that when a player has

[43] As self-attribution is closely linked to overconfidence, it comes as no surprise that it has been linked as the explanation to several phenomena that have been connected to overconfidence as well. For instance, Daniel et al. (1998) argue that self-attribution explains the excessive trading behavior observed in the markets.

[44] We, the authors, bitterly concede to have succumbed to systematic planning fallacy throughout the process of writing the thesis. Hence, we stand as a testimony to the cognitive bias of optimism.

made several baskets in a row, they are ‘*hot*’, and more likely to score on the next opportunity they have. However, no empirical evidence supports this claim. Therefore, the ‘*hot hand*’ phenomenon is an example of erroneous deduction about the population parameters on the basis of too few data points. The second rule of interest is commonly referred to as conservatism. Conservatism can be viewed as the opposite of law of small numbers. Under conservatism, agents tend to overweight the base rate at the cost of the sample rate. In other words, agents adapt too slowly to changes (e.g., Edwards, 1968). Intuitively one might think that the two cancel out on aggregate. However, that is not the case, and the agent is not Bayesian on average after taking into account both heuristics (Camerer, 1995). In fact, the saliency of the underlying model to which the data is being matched to determines the dominant effect: if the data is representative of a salient model then the sample rate is overweighed, if not, then the base rate is[45].

A closely related bias to conservatism is anchoring. When forming an estimate, agents anchor to initial - potentially arbitrary - value and then adjust away from it. However, the adjustments are not sufficient (Kahneman and Tversky, 1974). Indeed, in the case of conservatism, an argument can be made that agents anchor to the base rate and adjust then away from it. However the adjustments are insufficient, hence leading to the overweighing of the base rate (e.g., Ritter, 2003). In financial context, anchoring is observed in most speculative markets where the prevailing price is taken as the equilibrium price, the ‘*fair price’*. However, frequently the prevailing price significantly deviates from the fundamental value. Therefore, the belief that the prevailing value is the fair price is false (Mullainathan and Thaler, 2000). One potential explanation for anchoring is the excessive deliberation cost of computation to derive the fundamental price (e.g., Stracca, 2004). Therefore, to ease decision making, agents anchor on to representative values - sometimes completely arbitrary ones.

Belief perseverance and confirmation bias

Lord et al. (1979) documented that once agents have formed opinions, they are reluctant to search for evidence that would contradict their initial opinions. Furthermore, if such evidence is presented, agents will treat it with excessive skepticism. In fact, Rabin and Schrag (1999)

[45] Mullainathan (2001) provides a formal model to reconcile the evidence on underweighting and overweighting sample information.

argue that agents will go as far as interpreting spurious relationships in favor of their initial hypothesis in an attempt to avoid emotional cost of being wrong[46].

An extreme form of belief perseverance is the confirmation bias where agents misinterpret evidence contradicting the original hypothesis to be in fact in favor of it. An interesting example in the field of academic finance is the pivotal event study of Fama et al. (1969). The researchers concluded that their results were in line with, at that time, the nascent efficient market hypothesis. However, scholars have later on raised differing views on the interpretation of the results, arguing that Fama et al. (1969) ignored the clear price drift in the data: a clear sign of an inefficient market instead of an efficient one (e.g., Arbit and Boldt, 1984; Lee and Yen, 2008).

Loss and ambiguity aversion

Cognitive abilities have developed over a long period of time. In a hostile living environment, quite different from the modern environment, it was imperative to avoid unnecessary risks that could potentially be lethal. Risk aversion, on the other hand, causes negative events to have a greater impact on agents than positive events (e.g., Baumeister, 2001). Indeed, agents will go to great lengths to avoid materialization of negative events in order to protect themselves from emotional agony (e.g., Shefrin and Statman, 1985; Odean, 1998a). The described behavior is termed the disposition effect. Disposition effect, in the context of finance, hypothesizes that agents will hold losing assets longer than they should in order to avoid the anguish of realizing losses.[47]

Connected with risk aversion, and the resulting loss aversion, is ambiguity aversion. Agents dislike uncertain situations. In the context of modern finance, when agents do not know the objective probability distribution; hence, forced to use subjective probability distribution, agents will express a view on the probability as if the subjective probability distribution were objective. As subjective expected utility (SEU) does not allow agents to express their confidence on the subjective probability distribution, it cannot capture such aversion (e.g., Barberis and Thaler, 2003). Heath and Tversky (1991) argue that ambiguity aversion has much to do with how competent agents feel about assessing the subjective probability

[46] Thaler (2000) dubs the phenomenon as '*curse of knowledge*', a form of cognitive dissonance: when we know something we cannot imagine ever thinking otherwise - we are hence subject to belief perseverance.

[47] Ritter (2003) suggests that one of the manifestations of the behavior resulting from the disposition effect can be seen in the overall trading volumes of the aggregate stock market. In bull markets, agents sell more to realize small gains, hence pushing trading levels up. In bear markets, agents refrain from selling in order to avoid realizing losses. The aforementioned, as Ritter (2003) states, results in large systematic risk for broker houses.

distributions: the more confident agents feel about their competence, the more likely they are to take the uncertain choice. The manifestations of ambiguity aversion have been widely documented.[48] A finance-related example would is the tendency of investors to allocate their assets to familiar investments - for instance, the so called *'home bias'* can be seen to result from such behavior.[49] However, as we pointed out in the context of overconfidence, ambiguity is just one of the possible explanations offered to explain the tendency of agents to allocate their assets to familiar investments.

Framing: Procrastination and Mental Accounting

According to normative theories, framing of a problem should not impact agents' behavior as choices should be independent of the problem description. However, large body of evidence suggests that the way a problem is presented affects agents' behavior (e.g., Tversky and Thaler, 1990; Starmer, 2000). For instance, psychologists have documented that doctors make different recommendations when presented with survival probabilities instead of mortality rates, even though the latter is the complement of the former (Ritter, 2003).[50]

An important feature of framing is the concept of narrow framing. In narrow framing, agents limit the scope of the problem, presumably in an attempt to optimize deliberation costs under limited cognitive capabilities (e.g., Thaler, 1980; Read et al., 1999). The variable limited can differ from situation to situation, for instance, limiting the time frame of the problem is a common example of narrow framing coined procrastination.[51] Under procrastination, agents make rational choices at intervals that are irrationally short. Quitting smoking is a typical example of procrastination. Agents limit the time frame of the problem to one day, weighing the torment of withdrawal symptoms, with other cons, against the pros: needless to say that under one day time frame the choice is clear, and continuing smoking maximizes the agents' utility. However, after running through the same maximization on a daily basis for ten years, with the same '*rational'* choice, agents have continued smoking for ten years: obviously such behavior can have calamitous consequences for agents' health. Consequently, under procrastination, agents exhibit a strong tendency to overweight short-term utility at the expense of mid- to long-term utility resulting in sub-optimal choices under longer time

[48] For a thorough review of ambiguity aversion, see, for instance, Camerer and Weber (1992).

[49] See, for instance, French and Poterba, 1991; Lewis, 1999; Grinblatt and Keloharju, 2001; Huberman, 2001.

[50] In the field of economics, the phenomenon known as 'money illusion', first discussed by Fisher (1928), where low inflation is taken as zero inflation, has been suggested as an example of framing. For more discussion on the topic, see, for instance, Modigliani and Cohn, 1979; Shafir et al., 1997; Ritter and Warr, 2002; Ritter, 2003.

[51] For procrastination literature, see, for instance, O'Donoghue and Rabin, 1999a, 1999b, 2001; Brocas and Carillo, 2000; Fischer, 2001

horizon. Still, such behavior can be rational under expected utility maximization in the case of extreme constant discount rates. However, the constant discount rate axiom of expected utility maximization has been found faulty in several occasions (e.g., Frederick, Lowenstein and O'Donoghue, 2002). In fact, a time inconsistent model of discounting: '*hyperbolic discounting*', has been linked with procrastination. Under hyperbolic discounting, agents' impatience is steeper for near-term trade-offs than for long-term trade-offs. For instance, when asked whether or not an agent would like to have 100€ a year from now, or 50€ now, most agents will choose 50€ now. However, when asked whether or not an agent would prefer to have 100€ after six years, or 50€ after five years, most agents will choose 100€. Hence, agents exhibit diminishing sensitivity to time.[52]

Another common occurrence of narrow framing is mental accounting. In mental accounting, agents divide a problem into smaller pieces and then evaluate them in isolation from the larger - original - problem (e.g., Thaler, 2000; Barberis and Thaler, 2003). For instance, when agents are confronted with the problem of optimizing their asset portfolio, they segregate asset classes into separate mental accounts: for example, one account is for retirement money, one is of liquid assets and one is for risky investments. Once segregated, agents evaluate each mental account in isolation from the other accounts; hence, ignoring covariance between accounts - irrational behavior from the standpoint of modern portfolio theory[53,54].

Limited attention

As we have previously briefly discussed, agents are constrained by limited cognitive capabilities and are therefore distracted by irrelevant stimuli. For instance, the famous Stroop task (1935) asks subjects to name the color in which a word is printed. When the word does not match its print color (i.e., word *'blue'* is printed in red), subjects take more time to name the color. Another example is the phenomenon coined *'selective attention'* where subjects focus on a set of stimuli leading them to ignore other important stimuli. Examples of selective attention are dichotic listening[55] (e.g., Cherry, 1953; Broadbent, 1958; Moray, 1959) and

[52] For literature on hyperbolic discounting, see, for instance, Akerlof, 1991; Laibson, 1997; Caillaud and Jullien, 2000.

[53] Consequently, behavioral finance scholars have proposed a behavioral portfolio theory which encompasses several portfolio theories in place of one (e.g., Shefrin and Statman, 1994, 2000; Statman, 1999; Barberis and Huang, 2001).

[54] For more illustrative examples of mental accounting, for example: mental accounting and the importance of dividend policy, we refer the reader to Barberis and Thaler (2003).

[55] Subjects listen to two messages simultaneously but will only assimilate information from the message they are asked to follow — even the most elementary details are lost concerning the other message (i.e., language of the message).

inattentional blindness[56] (Simons and Levin, 1997). Also, divided attention in multiple tasks has been demonstrated to deteriorate subjects' performance on tasks (e.g., McLeod, 1977; Pashler and Johnston, 1998). Finally, the phenomenon dubbed *'cue competition'* shows that when subjects are asked to forecast stochastic variables based on multiple cues, the presence of irrelevant cues causes subjects to use less relevant cues (e.g., Baker et al., 1993; Busemeyer et al., 1993; Kruschke and Johansen, 1999).

Indeed, depending on the quality and quantity of information, it can take considerable time for agents to process information. Therefore, agents attempt to optimize the decision outcomes with the deliberation costs. However, in doing so, agents demonstrate a tendency to focus on salient information at the cost of more abstract information. Furthermore, agents take information as given without attempting to adjust it (e.g., Hirshleifer and Hong, 2002; Stracca, 2004; Arruñada, 2008; Barber and Odean, 2008). Consequently, agents are particularly vulnerable to manipulation and fads (e.g., Daniel et al., 2002, Stracca, 2004; Shiller, 2000b). In fact, companies can take advantage of agents' credulity by disclosing positive events in a salient manner and masking negative events with abstract disclosure (e.g., Skinner, 1994; Klibanoff et al., 1999). Furthermore, companies can overwhelm agents by disclosing more information in an attempt to make the disclosure even more opaque to agents. In other words, agents are in risk of losing the forest for the trees, when the quantity of information to be processed increases rapidly (Stracca, 2004).

In summary, limited attention predicts that in financial context, multiple sources of new information appearing simultaneously force agents to divide their attention resulting in gradual diffusion of information. Therefore, financial metrics exhibit underreaction to new information (e.g., Hirsleifer and Teoh, 2003, 2005; Hirsleifer, 2009).Opposing arguments do exist. For one, an argument can be made that agents can adjust rationally to the cognitive constraints they face by focusing on the important signals. Nevertheless, as described above, agents tend to focus on salient signals that are not necessarily the most vital ones. Indeed, Hirsleifer (2009) refutes the aforementioned argument on the basis that agents cannot determine ex-ante the importance of signals. He continues by arguing that even if agents were able to segregate important signals from less important ones, irrelevant stimuli would still affect information processing (e.g., Stroop task evidence). In fact, even the processing speed

[56] Subjects fail to perceive even the most extreme task-unrelated stimuli such as a woman in gorilla suit beating her chest in a supermarket (Simons and Chabris, 1999).

of sentiment-detecting algorithm that we describe in this study would be negatively impacted if more information was added.

2.2.2 Prospect theory

To settle the discrepancy between empirical evidence and the expected utility framework, a group of theories, labeled the non-expected utility theories, have risen to the challenge. Prospect theory, argued by many to be the most promising alternative (e.g., Camerer, 1998; Barberis and Thaler, 2003; Stracca, 2004, will be described next.[57]

The prospect theory is a truly descriptive theory with no ambitions in predicting agents' behavior. Indeed, prospect theory simply aims at capturing agents' attitudes towards risky gambles as parsimoniously as possible while striving to be analytically tractable (Barberis and Thaler, 2003).In contrast to expected utility theory, prospect theory evaluates gains and losses instead of the absolute level of wealth, an idea first proposed by Markowitz (1952). To motivate the focus on gains and losses, Kahneman and Tversky (1979) offer a compelling example of expected utility violation.

During a study, subjects were asked:

> *In addition to whatever you own, you have been given 1,000. Now choose between:*
>
> *A = (1,000; 50% probability)*
> *B = (500; 100% probability)*

The majority of subjects chose B over A. Subjects were then asked:

> *In addition to whatever you own, you have been given 2,000. Now choose between:*
>
> *C = (-1,000; 50% probability)*
> *D = (-500; 100% probability)*

One would expect consistent behavior from agents, as the two situations have similar final absolute wealth positions. However, on the second question, subjects chose C over D. Hence, subjects demonstrated seemingly irrational behavior under the expected utility theory. As a result, Kahneman and Tversky (1979) assert that in addition to focusing on gains and losses

[57] For other non-expected utility theories, we recommend the reader to get acquainted with: the weighted utility theory (Chew, 1983), implicit expected utility (Dekel, 1986; Chew, 1989), disappointment aversion (Gul, 1991), regret theory (Bell, 1982; Loomes and Sugden, 1982) and rank-dependent utility theories (Quiggin, 1982; Yaari, 1987; Segal, 1987, 1989).

over absolute wealth positions, agents are risk averse in the domain of gains and risk seeking in the domain of losses. In fact, Kahneman and Tversky (1979) assert that agents' value function is concave in the domain of gains and convex in the domain of losses. Also, the value function has a kink in the origin, illustrating loss aversion: a greater sensitivity to losses than to gains - a characteristic the reader might have observed when reading through the various cognitive biases influencing agents' behavior. Figure 1 demonstrates the hypothetical value function.

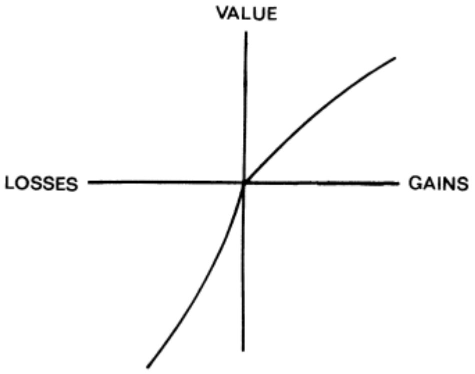


Figure 1: A hypothetical value function (Kahneman and Tversky, 1979)

The final key element underlining the prospect theory is nonlinear probability transformation. Kahneman and Tversky (1979) demonstrate that extreme probabilities are overweighed: in other words, agents' weighing coefficient does not behave in a linear manner consistent with the expected utility theory.

An important feature of prospect theory is that it can incorporate framing; particularly of interest is narrow framing, especially mental accounting. Due to the fact that the value function is non-linear, if agents segregate events that should be rationally pooled, the value of the pooled events can differ from the value of the segregated events. For instance, Shefrin and Statman (1984) present an interesting example in the context of corporate dividend policy that can be argued to be explained by prospect theory and mental accounting. Consider corporation A that plans on returning capital of 10€ to its shareholders. If A pays no dividends, the return will be in the form of a capital gain interpreted by agents as v(10). However, if A decides to return, for instance, 2€ as a dividend, the return can be coded by agents as: v(2) + v(8), or v(10), depending on whether or not agents are influenced by narrow framing bias known as mental accounting. What is interesting is, that under prospect theory, in the concave domain of gains, v(2) + v(8) > v(10), hence agents will prefer dividend and capital gain combination vis-à-vis a pure play capital gain - a crucial difference in preference when compared to the expected utility.

In their influential paper of 1992, Tversky and Kahneman generalize prospect theory to accommodate more than two non-zero outcomes. The successor has been christened the cumulative prospect theory (Starmer and Sugden, 1989; Tversky and Kahneman, 1992) and holds great promise as an alternative to the expected utility framework with much stronger psychological footing, yet maintaining analytical tractability.

While the list of empirical evidence contradicting the modern finance paradigm continues to grow, behavioral finance is still far from replacing the mainstream approach in its current form (e.g., Ritter, 2003; Stracca, 2004). For example, Shleifer (2000) maintains that behavioral finance has not yet reached the level of maturity which would allow it to provide a coherent unified theory of human behavior in the market context the same way expected utility and mainstream economics and finance have done. Nonetheless, several scholars (e.g., Thaler, 1999; Barberis and Thaler, 2003; Stracca, 2004) argue that behavioral finance will do so in the future.

Not all behavioral finance scholars share the rosy view concerning the future of behavioral finance. Some feel that behavioral finance will increasingly become part of modern finance (e.g., Ritter, 2003; Frankfurter and McGoun, 2002). Indeed, Frankfurter and McGoun (2002) suggest that the failure of behavioral finance to emerge as the new paradigm for finance is in fact paradoxically vindicating the teachings of behavioral finance. To elaborate, Frankfurter and McGoun (2002) argue that the same cognitive biases documented by psychologists, and advocated by behavioral scholars, are preventing behavioral finance paradigm from replacing the mainstream paradigm. In Frankfurter and McGoun’s view, the only hope for behavioral finance is the saturation of research that will force the academic society to seek new theories in an attempt to justify its own existence. Frankfurter and McGoun state that such development has already taken place as is demonstrated by the recent behavioral finance studies published in respectable publications, as well as the inspiring milestone of Daniel Kahneman’s 2002 Nobel Prize on the development of prospect theory.

In conclusion, the altercation between modern finance supporters and behavioral finance advocates is on-going, and unresolved. Empirical evidence exists on both sides to support the claims of both parties. Future research will determine the direction the debate will take, and seminal methodological articles such as Petersen (2009) can potentially shed more light on the results of the past. We remain hopeful for a resolution, and to paraphrase Richard Roll, for a theory and model that can better our understanding of asset prices.

2.3 Content analysis

Content analysis[58] in the field of finance is a relatively new branch of research. The underlying idea is to quantify investor sentiment from qualitative texts in order to be able to better, and quicker, analyze the information available in the markets. The motivation behind the studies has been to better explain the variation in several key financial metrics. For instance, content analysis seeks to enhance our limited understanding of variations in equity returns.

The prior literature in the nascent field has focused on identifying different methodologies for the estimation of investor sentiment, and from there on to study the impact of the sentiment on financial metrics. The key financial metrics used in previous studies have included: raw returns, abnormal returns, trading volume, return volatility and earnings. With promising new findings, content analysis has been gaining momentum in recent influential publications, and interest toward the field has increased. Furthermore, according to Demers and Vega (2010), there has been a surge in demand for firms selling data processed by linguistic algorithms (e.g. Ravenpack) implying that investment firms are waking up to the possibility of exploiting content analysis techniques in an attempt to reap returns.[59]

We aim to introduce the reader to the most influential findings in the field while giving insight to the theory behind the hypothesized relationship between investor sentiment and financial metrics. Furthermore, we offer a brief introduction to the content analysis process and methodology in general.[60]

The section will proceed as follow: first, we will briefly acquaint the reader with content analysis methodology; second, we will discuss the theory behind the hypothesized relationship between investor sentiment and financial metrics that will set the ground for our

[58] Also referred to as: natural language processing, information retrieval, computational linguistics, etc.

[59] Engelberg (2008) also states that hedge funds have become increasingly interested in using content analysis techniques to process textual data, and to trade on the extracted sentiment.

[60] For the interested reader, content analysis — outside the domain of finance — is an old field of study with its roots dating back to the 17th century when hymns were examined for word choices that threatened particular religious groups (Dovring, 1954). For an overview of the early research in content analysis, we refer the reader to Stone et al. (1966).

coming hypotheses in Section 3; third, we will discuss the key findings of content analysis studies in the domain of finance, their importance and the used methodologies.

2.3.1 Content analysis methodologies

Recent literature has used various approaches when turning text to quantified metrics. The idea is typically to transform qualitative information into a sentiment score – usually authors wish to distinguish positive vs. negative sentiment, and possibly measure the magnitude. As a criterion to compare whether the estimated sentiment correctly reflects the quantified, Mitra and Mitra (2010) suggest comparing a computer's annotations to how a human, or a human group (in particular a group of experts), interpret the text. Alternatively, one could also use market based measures, defining the sentiment after a market reaction (sentiment = market reaction). The aforesaid approach inherently assumes that the market reacts to a news story. Therefore, changes in a correctly constructed sentiment should correlate with stock performance metrics, such as returns, volatility or volume.

Sentiment analysis process

To analyze sentiment, one must first collect texts to process and to analyze them in order to construct a sentiment score. Mitra and Mitra (2010) split the information flow into information gathering (mainstream news, pre-news and web2.0/social media), pre-analysis, classification and assignment of sentiment scores, and analysis (vs. financial market data). Once completed, the analysis results can be fed into various quant models for return prediction, trading decisions or to assess risk. This approach is illustrated in Figure 2.

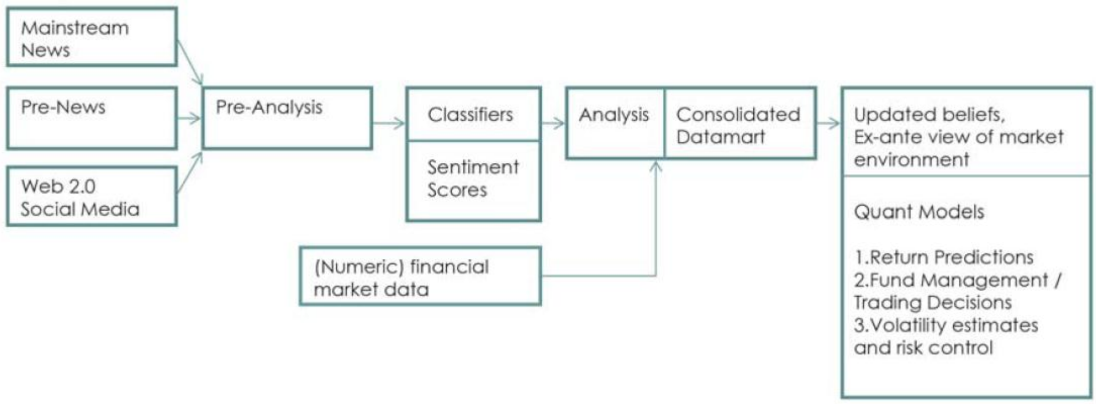

Mainstream News
Pre-News
Web 2.0 Social Media
Pre-Analysis
Classifiers
Sentiment Scores
Analysis
Consolidated Datamart
(Numeric) financial market data
Updated beliefs, Ex-ante view of market environment
Quant Models
1.Return Predictions
2.Fund Management / Trading Decisions
3.Volatility estimates and risk control

Figure 2: Information flow and computational architecture (Mitra and Mitra, 2010)

Various sources of information have been used in previous literature. Many authors (e.g. Engelberg, 2008; Li, 2006; Loughran and McDonald, 2011 have focused on 10-K reports that are easy to access and analyze as they follow a certain structure and relate by definition to a certain company. Examples of other media varies from general news (e.g. Antweiler and Frank, 2006; Tetlock, 2008) to analyzing message board posts (Antweiler and Frank, 2002).

Pre-analysis of information starts with the decisions on how media will be filtered from a database. For 10-K reports, the decision of which report relates to which company is straightforward. For news and social media, this needs to be carefully defined. News can also mention multiple companies and do not always relate to one company. Today’s news databases usually include some search functionality for a company ticker which is often the result of a machine learning algorithm by the database company. While Tetlock et al. (2008) choose to search news by the official company name and filter their results then further to ensure the news are highly relevant to the company, Engelberg (2008) relies on Factiva’s automatic classification of news by company code.

After sorting out the right companies, most authors also perform some pre-processing of the texts. This is necessary to, for example, include the heading to be a part of the text, or to deal with texts including elements not in a story-format, such as tables, pictures and disclaimers that could add unnecessary noise to a sentiment score.

Mitra and Mitra (2010) also recognize that it could be beneficial to identify stories that are current: news that report other old news are not so relevant anymore, as the information is not novel, and should often be given less weight or excluded from sentiment score metrics. Also, adjustments depending on news flow timing could be used. News flows have seasonality in them: at some points of day, week, month and year more news (and new information) come to the market than others. Finally, analyzing links between news should be considered, as news items often include a number of topics (e.g. a company’s earnings announcement will bring a wide variety of information to the table on different topics).

After preprocessing, news are classified to construct a sentiment score. Das (2010), also cited by Mitra and Mitra (2010), has identified six methods for classifying sentences: the naïve classifier and variations of the naïve classifier: the discriminant based classifier and the adjective-adverb phrase classifier; algorithms that determine the class based on the composition of lexicon items in a sentence: vector distance classifier and Bayesian classifier; and support vector machine (SVM). Das (2010) also proposes to use a voting scheme after using the number of classifiers, so that a message is given the category to which most classifiers would rank it.

Sentiment classifiers

The Naïve classifier (also known as "*word count*" and "*bag of words method*") works by counting the number of word occurrences, and assigns a label to the text based on what category of words are most common (e.g. positive or negative, or neural if no majority exists). To work, this method requires a lexicon, i.e. list of words that have been categorized as "positive", "negative", etc. Due to the ease of implementation, this is the most commonly used classifier and has been used in most studies in the finance domain (e.g. by Tetlock, 2008; with some additions by Engelberg, 2008; and Loughran and McDonald, 2011). As a modification of the naïve classifier, Das proposes a discriminant based classifier that assigns different weights to different words (e.g. 0.5 negative weight for a slightly negative word, and 2 for a highly negative word). The Adjective-adverb phrase classifier works also similarly to the naïve classifier, but considers only noun phrases that include adjectives or adverbs: e.g. "a strong profit" would be considered for classification, but "a profit" would not be included even if the word profit would exist in the lexicon.In addition, Engelberg (2008) experiments by adding the impact of simple negations that change the meaning of expressions (for example "not bad" vs. "bad").

The vector distance classifier assigns all words in lexicon as dimensions in vector space, and then describing each message as a vector. A training set of messages are pre-classified, and new messages are assigned polarity with vectors that have the smallest angle. Bayesian classifier, on the other hand, determines the count of each lexical item (e.g. a word) in a message. From a training set, it is possible to know with what likelihood each lexical item appears in a certain category. From word based frequencies, it is possible to calculate the probability that a message falls into a certain category, and assign the category with the highest probability to the message. For example O'Hare et al. (2009) uses multinomial naïve Bayesian classifiers to recognize sentiment in financial blogs.

Support Vector Machines (SVMs) are a classifier technique that is similar to cluster analysis but can be used in very high-dimensional spaces. Given a large number of texts and a training corpus, the SVM can classify texts, for example all words in the lexicon dimensions, and then clustering the texts based on information in the training corpus[61]. For example, this could be used to first identify which words are typically present in a positive sentence, and then to classify further sentences based on this. The advantage of SVMs would be their flexibility in

[61] For a more technical description of SVMs, see Das, 2010 and Vapnik and Lerner (1963); Vapnik and Chervonenkis (1964); Vapnik (1995); and Smola and Scholkopf (1998)

being able to learn features also in highly sophisticated environments. SVMs are used by e.g. O'Hare et al. (2009) to classify financial text.

Going further in sophistication with sentiment detection, Moilanen et al. (2010) have developed a method called "*quasi-compositional sentiment sequencing*" that we also use as the basis for our methodology. Compared to a base case word count, this method assumes that having polarities in different sequences can create a different outcome for the polarity of the whole sentence. For example, having a sentence with three words – "positive-negative-positive" – could be labeled as positive by majority vote with a word count algorithm. The logic of quasi-compositional sentiment sequencing, on the other hand, would be to ask "what kind of polarities have human annotators given to sentences that have words in the sequence 'positive-negative-positive'". To simplify sentences, the method compresses similar polarities together in the sequence (e.g. sentences with "positive-neutral-neutral-positive-positive" and "positive-neutral-neutral-neutral-positive" would be compressed to "positive-neutral-positive"). With this compression, the training sets required reduces significantly. For implementing the actual classification, the authors use a standard SVM approach and a readily annotated corpus (MPQA). Looking at the results of quasi-compositional sequencing, especially sentences with many different polarities (the authors use positive, negative, neutral and reversal) yield better results than simpler methods.

Considerations on classifiers

To work, classifiers often need supplementary databases: a dictionary includes the information of word categories (is a word an adverb, an adjective, a noun, etc.), a lexicon assigns words to various polarities (e.g. a list of positive words), and a training corpus of base messages shows examples of how different sentences should be classified. The contents of these databases can also vary significantly between studies, and e.g. changing from a general lexicon to a domain specific lexicon can make a large difference (see e.g. Loughran and McDonald, 2011). The most commonly used lexicon in the financial literature has been so far the General Inquirer's Harvard-IV-4 psychological dictionary (e.g. Tetlock et al., 2008; Engelberg, 2008).

In addition to sentiments, a sentiment algorithm can consider the window where the sentiment is detected, and also the magnitude of the sentiment. While most papers either consider sentiment on a document level or always categorize sentences, there are also other options for labeling a text with a certain polarity. O'Hare et al. (2009) introduce a concept of word (and sentence and paragraph) windows: they consider for the sentiment on a certain topic only text that has a distance of n to a topic word (e.g. only 5 words before and 5 words after a certain topic word).

For a human reader, it is also evident that the context of an expression impacts how strong the polarity should be. For example, Engelberg (2008) relates the negative sentiment on a sentence level further to one of six themes (positive fundamentals, negative fundamentals, future, outlook, environment, operations, and other) and identifies that they can be used to refine the perceived impact of sentiment.

Once classified, detected polarized words and sentences need to be combined to arrive at an aggregate sentiment score; in other words, "*sentiment of the day*". Authors have adopted various approaches for aggregation: e.g. Tetlock et al. (2008) combines all news of a particular day into one article and calculates the proportion of negative words in this article. On the other hand, Das (2010) labels each message as a "buy" or a "sell" signal, and then calculated the number of total buys and sells per day. The chosen approach has an impact: for example, a long article would typically have a larger weight with Tetlock's approach, whereas with Das's approach the weight of each article would be the same.

2.3.2 Stock metrics and investor sentiment: proposed link

During the 1980's, interest towards qualitative information began to surge. Empirical evidence that seemed to explain movement in financial metrics without any apparent change in quantitative information was growing (e.g., Shiller, 1981).[62] Confronted with the growing amount of apparent anomalies, scholars started to seek answers in qualitative texts leading some researchers to suggest that qualitative information could have incremental value above and beyond quantitative information in relation with financial metrics: e.g., equity returns (e.g., Roll, 1988; Cutler et al., 1989).[63]

Information content in financial texts

Concurrently with the increasing interest in the informational content of qualitative text, research was documenting managers' tendency to voluntary disclose information to investors in order to align investors' expectations of future performance with management's own

[62] For discussion on the anomalies, see Section 2.1.

[63] In spite of the suggested link between qualitative information and asset prices, Cutler et al. (1989) find no link between important qualitative news unaccompanied by quantitative macroeconomic events and market returns.

assessment (e.g., Ajinkya and Gift, 1984; Hassel and Jennings, 1986; King et al., 1990).[64] In 1994, in his seminal article using a sample of voluntary earnings disclosures for 93 NASDAQ firms, Skinner (1994) showed that managers tended to disclose good news in point estimates and bad news in qualitative text. In fact, according to Skinner, previous literature had forgone an important source of information when neglecting to take into account the impact of qualitative information on financial metrics. Sloan (1996) agrees with Skinner by positing that the power of market efficiency tests can be improved if the strategic nature of published disclosures can be exploited. Later on, researchers have agreed with Sloan, suggesting that qualitative information provides an interesting opportunity to improve tests on market efficiency (e.g., Antweiler and Frank, 2006; Li, 2006; Davis et al., 2008).

In line with the forthcoming findings of Skinner, Subramanian et al. (1993) shows that annual reports of profitable firms are significantly easier to read than those of poor performers — implying that poor performers disclose information in a more complex manner using lengthier and more difficult qualitative texts. Later on, Li (2008) demonstrates with a sample of 10-ks that managers attempt to hide adverse information through less transparent disclosure via qualitative text. Li (2008) employs a fog index in combination with the length of a document to measure annual report readability.[65] The evidence supports Skinner's (1994) findings and is in line with Subramanian et al. (1993), showing that annual reports of firms with lower earnings are harder to read. Moreover, increase from last year's earnings will decrease the complexity of the annual report corresponding to the increased earnings period.

Besides containing information on past and contemporary negative fundamentals, qualitative text has been linked to forward looking estimates vis-à-vis the backward looking focus of quantitative point-estimates (e.g., Li, 2006, 2008). The rationale is that managers have more freedom in writing qualitative texts which are loosely regulated vis-à-vis quantitative information which is strictly regulated (e.g., Li, 2006; Davis et al., 2008).[66] Therefore, managers are more inclined to disclose future estimates using qualitative information. As a result, information extracted from qualitative texts can in fact have incremental value above

[64] For more recent findings on the topic, we refer the reader to, for instance: Verrecchia (2001).

[65] Fog index combines the number of word per sentence and the number of syllables per word to create a measure of readability while the length of the document proxies for higher amount of information processing required by the reader.

[66]SEC does not address language use in earnings press releases per se (Trautmann and Hamilton, 2003). However, there are some guidelines for qualitative text disclosure: in 1998 SEC issued a new plain English disclosure guidelines that, for instance, prohibited double negatives, legal jargon and highly technical business terms (Glassman, 2005).

and beyond quantitative information when predicting financial metrics. Dye and Sridhar (2004) posit that hard and soft information should be taken into account in tandem. Also in Engelberg's (2008) view, finance scholars should focus their attention in analyzing the heterogeneity of information: soft vs. hard, as some corporate finance scholars have done (e.g., Stein, 2002; Petersen, 2004).[67]

Tetlock et al. (2008) offer two significant rationales for using qualitative texts: first, by analyzing all relevant news, researchers can analyze and judge the directional impact of a limitless number of events simultaneously through a proxy of investor sentiment. Furthermore, while examining all newsworthy events, researchers effectively limit their possibility for data dredging on a specific anomaly. Second, most investors receive their information secondhand Therefore qualitative texts can have incremental value over first hand quantitative information concerning a firm's fundamentals as they are better proxies for the information set that investors use.

Li (2006) goes even further by suggesting that the documented anomalies can in fact be proxies for a same undocumented omitted variable and as such the anomalies are not in fact independent of each other. Fama and French's (2006) findings lend some support for Li's proposed hypothesis: Fama and French find a strong correlation between the different variables associated with known anomalies. To study the hypothesis, Li regresses Fama and French's (2006) variables on his risk sentiment, and finds that several of the variables are significant and explain risk sentiment. Furthermore, Li suggests that the variables that are insignificant are insignificant due to multicollinearity issues arising from the strong correlations between the variables. In conclusion, Li suggests that sentiment based on qualitative texts can potentially offer a more accurate and independent test of market efficiency.[68]

Market impact of financial texts

Besides the informational content of qualitative text, the linguistic style of the text can have an impact on financial metrics even in the absence of new information (e.g., Davis et al.,

[67] Corporate finance literature has argued that soft information increases the cost of transmission.

[68] Loughran and McDonald (2011) take a more negative view concerning the status of sentiment, arguing that there is no link between sentiment and returns in the existing literature. In fact, sentiment is most likely a proxy for other contemporaneous information such as accounting numbers. However, Loughran admits that even though sentiment might not be a true driver for returns, it might still be an efficient way to capture other sources of returns.

2008; Henry, 2008). As framing is documented to have an impact on agents' behavior,[69] the linguistic style can impact agents' decisions even in the absence of new information. Indeed, several studies have found that recipient of messages are attentive to both content and style (e.g., Petty and Cacioppo, 1986; Chaiken, 1987; Kruglanski and Thompson, 1999; Chung and Pennebaker, 2007). For example, the way the stock market commentators describe price movements influences investors' expectations of future prices even in the absence of any fundamental reason for price movements (Morris et al., 2005).

If qualitative text has an impact on financial metrics, the question remains whether or not that impact is disseminated by the markets instantaneously as suggested by the EMH. As we have previously established, managers seem to be inclined to disclose negative news through qualitative text implying that qualitative text offers a better medium to communicate negative events for some reason. One such reason might be that qualitative text is harder for agents to process (e.g., Petersen, 2004; Li, 2006, 2008; Davis et al., 2008; Engelberg, 2008); therefore, managers disclose negative events through qualitative text to mask the event's full impact from agents (e.g., Bloomfield, 2002). Grossman and Stiglitz's (1980) *'incomplete revelation hypothesis'* supports such assertion, stating that information that is more costly to process is less completely reflected in market prices.

Also the limited attention theory[70] supports the slow incorporation of information from qualitative texts into financial metrics. The theory posits that agents have limited cognitive resources and hence possess limited capacity to allocate to information processing. As a result, when agents are forced to divide their attention among several information cues, information is incorporated into financial metrics with delay resulting in underreaction to new information. Indeed, there is a large corpus of studies documenting underreaction to different events across different firms (e.g., Hong et al., 2007; Hou, 2007; Cohen and Frazzini, 2008).[71] Limited attention offers an explanation to underreaction in the context of both quantitative and qualitative information.[72] However, in the case that qualitative information is more costly to processes, limited attention would predict that information based on qualitative text experiences greater underreaction as agents focus on the more salient information

[69] See Section 2.2.1. for discussion on framing bias.

[70] For more discussion on limited attention, see Section 2.2.1

[71] For more evidence, see Section 2.1.

[72] Li (2006) takes a more negative stance towards limited attention's ability to explain quantitative information anomalies, stating that they are well documented and easily exploitable. Therefore, require minimal processing and attention. However, Li agrees that limited attention should play a major role in the dissemination of qualitative information.

(quantitative information) at the cost of the more abstract information (qualitative information).[73] Indeed there is evidence that supports the claim: for example, prior literature has found that stock markets react to previously published news implying that relevant information is neglect at the time of the release (e.g., Ho and Michaely, 1988; Huberman and Regev, 2001).

Another potential explanation for underreaction to information contained in qualitative texts is the leniency in qualitative text regulation that can cause agents to discount information in qualitative texts (Davis et al., 2008). Mercer (2004) posits that as there are no third party auditors for qualitative text — as there are for quantitative numbers — agents will discount information from qualitative sources and therefore underreact to such information[74].

2.3.3 Sentiment and financial metrics: findings and methodologies

With the increasing interest towards informational content of qualitative texts, scholars have attempted to estimate investor sentiment (information content and tone) from different sources of qualitative texts in order to study the relationship between sentiment and important financial metrics. Major sources for qualitative texts include: news articles (e.g., Chan, 2003; Antweiler and Frank, 2006; Tetlock, 2007; Kothari et al., 2008; Tetlock et al., 2008; Bushee et al., 2010), company press releases[75] (e.g., Davis et al., 2008; Engelberg, 2008; Henry, 2008; Bhattacharya et al., 2009; Demers and Vega, 2010), 10-ks (e.g., Li, 2006, 2008; Kothari et al., 2008; Loughran and McDonald, 2011) and Internet message boards (Antweiler and Frank, 2004; Das and Chen, 2006). Recently, social media (e.g. Twitter) has also sparked the interest of researchers. Also, the impact of aggregate market news volume on financial metrics has been studied (e.g., Mitchell and Mulherin, 1994; Hirsleifer, 2009) as well as the impact of firm specific media coverage (e.g., Barber and Odean, 2008; Fang and Peress, 2009; Loukusa, 2011). Majority of the studies has tried to link investor sentiment with one of

[73] As shown in Section 2.2., with representativeness bias, the saliency of the model is the vital component in determining which information source agents overweight at the cost of the other. In financial context, point-estimates are often more salient due to valuation models utilizing point-estimates as their inputs. Therefore, qualitative text is often more abstract to agent, and is therefore underweighted in many cases.

[74] In connection with the aforementioned, Krishna and Morgan (2004) with Demers and Vega (2010) show that multiple experts (sources of information: i.e., news and analyst press releases) improve information credibility and the subsequent reaction to the information in qualitative text. As a result, it is possible that agents underreact to news as they discount the informational value of qualitative text due to loose regulation.

[75] Including various forms of corporate press releases and disclosures such as: earnings announcements, IPO prospectuses, voluntary disclosures, etc.

the following financial metrics: raw returns[76], abnormal returns[77], trading volume[78], return volatility[79] or earnings[80].

Sentiment and returns

Majority of studies[81] has found a significant reaction between investor sentiment changes and returns. The type of reaction found has been underreaction.

Chan (2003) finds underreaction to news with a comprehensive list of news publications including Dow Jones Newswire service, which Tetlock et al. (2008) find to have novel information content resulting in an underreaction to sentiment changes. Furthermore, studies on other qualitative text sources have found results consistent with underreaction (e.g., Li, 2006; Engelberg, 2008; Demers and Vega, 2010; Loughran and McDonald, 2011). Also, prior literature has found that simultaneously released information sources distract investors resulting in stronger underreaction to information (Hirsleifer et al., 2009). Hirsleifer et al., name this phenomenon as '*the distraction hypothesis'* which is based on limited attention theory. Tetlock et al. (2008) study offers some support for Hirsleifer et al.'s findings demonstrating evidence that earnings related news cluster around earnings announcements, while earnings announcements occur almost always during the same time period, and that news that have the word stem *'earn'* predict stronger underreaction. Tetlock et al. interpret that such news articles deal with fundamentals and therefore qualitative texts concerning fundamentals have more information content and subsequently more impact.[82] However, an alternative conclusion might be drawn to support Hirsleifer et al.'s later study, suggesting that news with the stem *'earn'* are news dealing with earnings announcements and as such are released in close proximity to earnings announcements. If this is the case, such news are released during a time when qualitative information, with quantitative information, floods the

[76] E.g., Chan, 2003; Antweiler and Frank, 2004; Tetlock, 2007; Tetlock et al., 2008.
[77] E.g., Chan, 2003; Li, 2006; Engelberg, 2008; Tetlock et al., 2008; Hirsleifer et al., 2009; Demers and Vega, 2010; Loughran and McDonald, 2011.
[78] E.g., Antweiler and Frank, 2004, 2006; Tetlock, 2007; Hirsleifer et al., 2009; Loughran and McDonald, 2011.
[79] E.g., Antweiler and Frank, 2004; Demers and Vega, 2010; Loughran and McDonald, 2011.
[80] E.g., Li, 2006, 2008; Tetlock et al., 2008; Loughran and McDonald, 2011
[81] Exceptions include of Antweiler and Frank (2004) and Das and Chen (2006). However, the aforementioned authors study internet message board data that does not seem to have incremental value over quantitative information. Furthermore, both studies have had fairly limited sample sizes. Also, their methods rely on more sophisticated - and complex - methodologies for estimating investor sentiment that have not yet yielded good results.
[82] Tetlock et al. (2008) do not discuss any alternative interpretations for their result, but conclude that the finding is in line with their initial hypothesis. Therefore, one might present the argument that the psychological bias of 'belief perseverance' is clouding their interpretation.

market causing greater distraction with more severe underreaction in line with the distraction hypothesis suggested by Hirsleifer et al., and the limited attention theory.

Some studies document also overreaction to sentiment changes: Tetlock (2007) finds that Down Jones Index raw returns overreact to Wall Street Journal [WSJ] articles within one day of the event and reverse back to fundamentals within the next 5 trading days. Tetlock infers that the news articles carry no additional information content, and the reversal is consistent with the EMH. However, Tetlock (2007) studies WSJ articles which are later found to have no significance in relation to financial metrics in a study by Tetlock et al. (2008). In fact, Tetlock et al. (2008) hypothesizes that WSJ articles simply recapitulate previous news and therefore have no new information content, and as such should have no reaction. However, in light of the framing bias discussed in previous sections, one can argue that the tone of WSJ articles affects agents' decisions resulting in overreaction to sentiment documented by Tetlock (2007), and the subsequent gradual reversal back to fundamentals. Therefore, depending on the information content of qualitative text, underreaction and overreaction are both possibilities.[83] Indeed, Antweiler and Frank (2006) also find initial overreaction to sentiment changes from WSJ articles that reverses later on. The findings of Antweiler and Frank give support to our reasoning on overreaction with sentiment changes based on WSJ articles.

In terms of event windows used in the prior literature, most of the studies have focused on short-term reactions with event windows equal to, or shorter than, 4-days (e.g., Tetlock, 2007; Davis et al., 2008; Tetlock et al., 2008; Loughran and McDonald, 2011).[84] As a result, the intermediate- and long-term effects of sentiment changes, and the impact of such results on the efficiency of the market, have been neglected — to some extent — by prior literature. Some recent studies have employed longer event windows (e.g., Engelberg, 2008; Demers and Vega, 2010)[85] in addition to short-term windows with findings that suggest that underreaction continues long after the opening of the event window in violation with the proposition of the EMH.

[83] Tetlock (2007) dubs the alternative theories as: sentiment hypothesis (reversal) and information hypothesis (underreaction).

[84] Potentially due to the critique that has been directed towards longer event windows and the consequent increase in the severity of 'bad model' problems associated with them (e.g., Fama, 1991, 1998; Malkiel, 2003) – more on Section 2.1.

[85] Engelberg (2008) states that his study is the first to show content of financial media can predict asset prices in a longer time horizon.

Besides studying the link between returns and sentiment, scholars have attempted to build trading strategies to test the economic significance of the relationship between sentiment and returns (e.g., Li, 2006; Tetlock, 2007; Engelberg, 2008; Tetlock et al., 2008; Loughran and McDonald, 2011). As some EMH proponents have argued, an economically more feasible form of the EMH (e.g., Jensen, 1978) is not against market efficiency but in fact in line with it. Therefore, for an anomaly to be against all the definitions of market efficiency, it must stand and survive the test of economic significance.

To study economic significance, trading strategies are constructed. Most of the trading strategies are based on prior year's distribution of sentiment. By utilizing prior year's distribution, the strategies take both long- and short-positions based on the magnitude of the sentiment change vis-à-vis prior year's sentiment distribution (e.g., Li, 2006; Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011). At the moment, it seems that the economically feasible form of the EMH stands unwavering against the results. In other words, the link between sentiment change and abnormal profits is not strong enough to survive transaction costs (e.g., Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011).[86] However, as Tetlock et al. (2008) point out, the methodologies used for estimating investor sentiment are rudimentary as of now and as such the estimates are biased downwards due to measurement error. Therefore, the magnitudes of the sentiment coefficients are in reality more significant, as would the subsequent results be with more accurate sentiment estimates. Also, Tetlock et al. (2008) suggests that by creating more elaborate trading strategies, the economic significance might turn out to be different, and could be in violation with the economically more feasible form of the EMH.

Sentiment and trading volume

Depending on the information content of qualitative texts, sentiment level changes can impact trading volume in differing ways. As Tetlock (2007) notes, sentiment changes with no new information content will result in overreaction and therefore a volume increase. However, in Tetlock's view, the hypothesis with sentiment change in the case of new information is unclear.[87]

[86] Some studies have been able to establish trading strategies that have been able to stand the test of transaction costs, and emerge as victors (e.g., Li, 2006; Engelberg, 2008)

[87] In line with his hypotheses, Tetlock (2007) finds that sentiment changes in WSJ articles are matched with overreaction leading to increase in trading volume.

Hirsleifer et al. (2009), state that in the presence of greater distraction, as measured by larger level of simultaneous earnings announcements, investors will underreact to new information and therefore volume will not be impacted, or it will be lower, in the short-term. Hirsleifer et al. interpretation suggests that underreaction is not accompanied by increase in volume in short-term. As we have demonstrated in the previous section, in the context of returns, most studies have found that qualitative texts have new information, and that information is incorporated into prices with delay: underreaction. Therefore, one might infer that in the case that information content dominates over style and tone of the text, underreaction dominates. Subsequent short-term volume changes would therefore not be abnormal, as suggested by Hirsleifer et al. findings regarding aggregate market earnings announcements and volume levels.

On the other hand, Loughran and McDonald (2011) find that abnormal trading volume increases with sentiment changes. Loughran and McDonald's results are more consistent with the interpretation that their qualitative information contains new content, and therefore the finding should not be attributed to the disagreement hypothesis[88] suggested by Tetlock (2007). However, Loughran and McDonald's event window differs from that of Tetlock's (2007) potentially casting some light to the difference. Indeed, Loughran and McDonald's event window is longer than that of Tetlock (2007); 4-days vis-à-vis 1-day event window. Therefore, Loughran and McDonald's results might be driven by the fact that volume levels reacts to new information with a lag.

We propose that the relationship between trading volume and sentiment seems to be related to information content vis-à-vis style and tone of qualitative texts as well as the event window length under review. Antweiler and Frank (2006) findings support our proposition: they find that initial overreaction is accompanied by increase in volume followed by declines in volume — such would be the reaction with texts dominated by tone over content. In cases where investors react more to new information content (over tone of text), sentiment changes should have either no reaction — or slightly positive reaction — with trading volume in the short-term followed by slight increase in volume in the long run.

[88] For details on the hypothesis, we refer the reader to Tetlock (2007).

Sentiment and volatility

The impact of sentiment on return volatility has not been studied extensively in prior literature. To our knowledge, only the recent studies of Demers and Vega (2010) and Loughran and McDonald (2011) have studied the relationship. If we go back further in time, we find that Antweiler and Frank (2004) have also studied the impact of sentiment on return volatility with a limited sample of internet message board posts.

All of the aforementioned studies find a relationship between return volatility and a change in investor sentiment. The relationship seems clear: changes in sentiment (both towards more positive and towards more negative sentiment) predict increases in future return volatility.

Sentiment methodologies used in finance research

Crucial part of the research concerning the impact of qualitative text on financial metrics is the process of quantifying that text into a sentiment score (e.g., Loughran and McDonald, 2011). As content analysis is a rather new field in the domain of finance, the methodologies used so far in the extraction of investor sentiment have been relatively simple and rudimentary. In fact, Tetlock et al. (2008) argues that more complex methods of content analysis face two significant drawbacks: first, the need for human judgment; second, the difficulty to replicate the study by later research. Instead, Tetlock et al. advocate the use of established dictionaries (i.e., Harvard Psychology Dictionary) with word counting (i.e., '*bag-of-words*' vector word counts) that give results four crucial attributes: parsimonious, objective, replicable and transparent. Tetlock et al. continue by arguing that the aforementioned attributes are crucial in the early stage of research in content analysis in the domain of finance. However, in slight contradiction with their own proposition, Tetlock et al. also urge researchers to develop less noisy measures of investor sentiment - effectively urging the use of more complex methodologies in estimating sentiment.[89]

Li (2009) rejects Tetlock et al. (2008) notion of using established dictionaries as there are no readily available dictionaries for the setting of business, and advocates also the use of statistical methods in word categorization. Such methods run head-on to the critique presented by Tetlock et al. (2008). Loughran and McDonald (2011) offer a potential solution to the dilemma by developing their own dictionaries for the domain of finance in the context of 10-k

[89] By arguing the following, we link more complex methods with better estimates of sentiment. However, that is not necessarily the case. Nevertheless, we do argue that at some point it is necessary to move towards more complex methods in estimating sentiment if we are to improve the accuracy of sentiment estimates as sentiment is extremely complex variable that requires a complex model to take into account all the required nuances.

reports.[90] In fact, Loughran and McDonald show that Harvard Psychology Dictionary misclassifies words more than 70% of the time in finance context. On top of adding noise to the estimate, Loughran and McDonald assert that the misclassifications can in fact result in spurious correlations and type I errors. Loughran and McDonald conclude that future research should use their dictionaries in finance context.

As can be inferred from above, word counts (i.e., '*bag-of-words'* vectors) with established dictionaries have been the most prominent method for quantifying qualitative texts in finance context (e.g., Tetlock et al., 2008). Yet, exceptions such as Antweiler and Frank (2004, 2006) and Das and Chen (2006) do exist that have used more complex methodology[91]. Also, more recently, Engelberg (2008) employs a typed dependency parsing method to extract sentiment using sentence structures. Engelberg states that his study is the first study in the domain of finance to employ content analysis methodology that utilizes sentence structures instead of words.

In order to employ word counts, researchers need to either choose dictionaries, or create their own. The early studies of sentiment in finance used only few specific words instead of full dictionaries. For instance, Li (2006) uses words *'risk'* and *'uncertainty'* to measure risk sentiment. Later on, since the seminal study of Tetlock (2007), academia moved on to use Harvard Psychology Dictionary, and to be more specific, the negative word category of the dictionary.[92] As Tetlock (2007) shows in his study, the fraction of negative words to total words is a good proxy for sentiment that performs equally well with a variable created from all the 77 different Harvard Psychology Dictionary categories.[93] However, with the recent influential paper of Loughran and McDonald (2011), the choice of dictionary for future research in the domain of finance is uncertain.

[90] One can argue that Loughran and McDonald (2011) also fall prey to the critique of subjectivity. However, the authors construct their dictionaries to be as holistic as possible to avoid such critique.

[91] Antweiler and Frank (2004) and Das and Chen (2006) fail to find significant results with their main variable of interest: returns. However, their failure can be also be a function of their sample: internet message boards, or sample size, rather than the methodology of choice.

[92] Tetlock (2007) suspects that negation increases noise with positive words and causes them to lose significance. However, Engelberg (2008) tests the amount of negation present in DJNS articles and finds that there is no substantial negation. Therefore, he rejects Tetlock (2007) explanation and suggests that misclassifications are responsible. Yet, when Loughran and McDonald (2011) create their own positive word list, they are unable to increase the significance of their results. Loughran and McDonald conclude that truly positive words are hard to isolate.

[93] The latter variable is created based on past year's data using principal component analysis which combines the 77 categories into a linear combination variable that captures maximum amount of variance out of the 77x77 covariance matrix.

2.4 Unexplored areas in literature

As content analysis is a relatively young field of study in the domain of finance, there are several different areas of interest left unexplored requiring research. We aim to quickly introduce the reader to the most significant areas of interest requiring future development and research.

As discussed in Section 2.3., used methodology for the estimation of investor sentiment stands in the center of all research in the area. Therefore, it comes as no surprise that there is plenty of work to be done in the area. As discussed in Section 2.3.2., there are differing views on the required complexity of methodology for sentiment estimation. However, as research progresses in the domain of finance, researchers must utilize more up-to-date methods from computer-science in order to build a more realistic model to better capture sentiment. Therefore, future research should focus on trying to leverage content analysis techniques developed by different disciplines in order to build a better methodology for estimating investor sentiment in financial context. The aforementioned is advocated by Tetlock et al. (2008) and Li (2009), among others.

Besides developing and leveraging new methodologies, existing methodologies require work. To paraphrase Berelson (1952): content analysis stands or falls by its categories. Therefore, while research continues to employ word counts based on dictionaries, the used dictionaries are in the center of the credibility and reliability of results. With the influential paper of Loughran and McDonald (2011), the status-quo position of Harvard Psychology Dictionary has come under scrutiny.[94] Indeed, Loughran and McDonald advocate the use of their dictionaries in the context of finance. However, Loughran and McDonald's dictionaries were created in the context of 10-k reports, and so far no other published study has tested Lougran and McDonald's dictionaries — to our knowledge. In fact, Loughran and McDonald posit that future research should test their dictionary using different qualitative text sources from 10-ks. Also, before scholars convert from using Harvard Psychology Dictionary into using finance specific dictionaries, the results of Loughran and McDonald (2011) should be replicated using a different sample to ensure that the results are accurate — even in the context of 10-ks.

In addition to methodological areas of interest, current studies have focused on studying short-term event windows without focus in the intermediate and long-term effects of

[94] To be fair, Loughran and McDonald (2011) do show that if employing a term weighting scheme, the introduced bias by Harvard Psychology Dictionary attenuates.

sentiment. Antweiler and Frank (2006) urge future research to include longer event windows in their studies to account for long-term effects of sentiment changes. Indeed, Engelberg (2008) and Demers and Vega (2010) have included such windows. However, there is yet much work to be done in analyzing sentiment change impact in different event windows.

Intertwined with the issue of event window lengths, is the impact of investor sentiment on market efficiency, and the current debate between behavioral finance and efficient market proponents discussed in sections 2.1. and 2.2. Indeed, several content analysis papers have urged future research to focus on the impact of sentiment on market efficiency (e.g., Antweiler and Frank, 2006; Li, 2006; Davis et al,. 2008; Engelberg, 2008; Tetlock et al., 2008). Therefore, studying the reaction of financial metrics on sentiment changes, and the subsequent theory development is crucial in the future.

In addition to the above, consolidating media research into a comprehensive model of media effect on financial metrics offers great potential. Firm specific media coverage effects (e.g., Fang and Peress, 2009), aggregate market wide news activity (e.g., Hirsleifer et al., 2009) and firm specific extracted investor sentiment scores (e.g., Tetlock, 2007; Tetlock et al. 2008; Loughran and McDonald, 2011), have all been linked with future performance of financial metrics. Therefore, future research should attempt to analyze the relationship of all the aforementioned variables on financial metrics simultaneously to create a holistic picture of the impact of media on financial metrics. Indeed, there is no evidence to suggest that the effects are mutually exclusive as of now.

Also, future research should attempt to incorporate all qualitative texts to truly analyze all the prevailing sources of sentiment simultaneously (as opposed to studying e.g. only the impact of an individual newspaper column). Indeed, some studies have used wide sources of qualitative texts at once when estimating sentiment (e.g., Chan, 2003; Kothari et al., 2008). Yet, the area offers great potential and should be of interest to future research.

Finally, employing proven techniques to different data sets from different countries can also prove to be interesting. The subsequent findings on the potential difference can also shed light to studies relating to the different informational efficiency of different markets (i.e., fringe markets). Furthermore, developing new dictionaries for different languages, and testing the differences between information dissemination between different language sources can be extremely interesting. However, these topics are most likely not yet topical, but offer potential in the future once the nascent field has matured.

3 HYPOTHESES AND CONTRIBUTION

In this section, we aim to highlight our place in the literature while describing the hypotheses we are studying. We will first describe the contribution our study makes to the field, and then move on to the hypotheses that underline our research.

3.1 Contribution

In section 2.4., we mapped out the various interesting areas requiring further research in the field of content analysis in finance. To contribute to the existing literature, we aim to study some of these areas. As a result, we contribute to the existing literature by:

- Introducing new categories to the used finance lexicon. Recent models have typically used word lists of positive and negative words, focusing especially on negative words. However, as pointed out by Loughran and McDonald (2011), there are also many other word categories that can be helpful in detecting semantic orientations in financial and economic texts. To complement these, we add to the finance lexicon directional verbs and financial entities - words that require a directional verb to receive a polarity.
- Testing the prior literature's principal methodology for estimating investor sentiment: the vector word count, with the two most prevalent dictionaries utilized: the Loughran and McDonald (2011) dictionary, and the Harvard Psychology dictionary, in order to clarify the efficacy of the extant methodology, and the preferred dictionary for the methodology.
- Testing Loughran and McDonald's (2011) dictionary in an out-of-sample test with qualitative texts other than 10-k reports.
- Developing a novel and a superior methodology, the Linear Phrase-Structure -model (LPS), for detecting semantic orientations in financial text Our methodology extends the categories used for determining sentiment and works beyond the level of detecting word lists[95].

[95] To benefit from the domain-specific knowledge, which we add into the finance lexicons, one needs to have a model that is not restricted to frequencies of positive or negative words but is able to take the entire phrase-structure into account. We extend the polarity-sequence framework of Moilanen et al. (2010) by accommodating

- Testing the three documented major media variables on financial metrics simultaneously in an attempt to create a holistic media model, and to isolate the significant media factors that drive variations in financial metrics.
- Analyzing the impact different media factors have on financial metrics in a robust study employing a comprehensive cross-section of different qualitative texts with multiple event windows in order to attenuate concerns over data dredging.
- Drawing conclusions on how limited attention (and related behavioral finance theories) could underlie the link between link between financial metrics and media factors.
- Putting forward evidence concerning the efficiency of the market and the informational content of qualitative texts.

3.2 Hypotheses

We will now describe our hypotheses for the study that are based on prior literature introduced in Section 2. We begin by describing our hypothesis for the performance of our novel sentiment estimation methodology (Linearized Phrase-Structure -model) vis-à-vis a word count method that employs the two most prominent existing dictionaries. We will then describe our hypotheses for the impact of our main variables on abnormal returns. From there on, we will describe the hypotheses relating to trading volume and our main variables. Finally, we will explain the hypotheses concerning abnormal volatility and our main variables. For each of the subsections, we describe first the hypothesis to link the variable to sentiment, then to market news volume, and finally to firm-specific news volume. After describing each hypothesis, we also outline what alternative explanations be possibly exist in the market in the case that our hypothesis would not hold.

3.2.1 Sentiment methodology hypotheses

At the core of our study, we have developed a new methodology for better measuring sentiment. Compared to '*bag-of-words methods*', we expect Linearized Phrase-Structure -model to lead to more accurate estimates of sentiment. Therefore, our first hypothesis is:

elements which are particularly relevant for the financial domain. The most important differences to this method that we use to enhance the method include: (1) the inclusion of finance-specific entities into the polarity sequence model; and (2) the inclusion of interactions between financial concepts and verbs or other direction-giving expressions

> ***H1:*** *Linearized Phrase-Structure -model outperforms the existing prevalent methodologies used for sentiment estimation in financial context.*

Alternatively, in the case that the bag-of-words method with prevalent dictionaries outperforms Linearized Phrase-Structure -model, we suggest that the outcome is because Linearized Phrase-Structure -model limitations lead to a significant measurement error.

3.2.2 Abnormal return hypotheses

In terms of returns, we aim to find a connection to sentiment. Our hypothesis stems from the limited attention theory, and states as follows:

> ***H2a****: Investor sentiment forecasts future abnormal returns in all event windows through underreaction to sentiment changes.*

This also implies that we expect information content to dominate over tone and style effects. On the other hand, we recognize the fact that tone and style effects could also dominate over information content. Also, if sentiment has no impact on abnormal returns, this could be due to the fact that markets efficiently incorporate new information to prices, and that at an aggregate market level markets are efficient, or that there is no new information in news.

In addition to sentiment, we expect aggregate market news volume to have an impact on returns. More precisely, we expect news volume to distract investors in line with limited attention theory, increasing the underreaction. Hence the hypothesis:

> ***H2b****: Aggregate market news volume has a positive relationship with abnormal returns in longer event windows, while having a neutral, or negative, relationship with the 1-day event window.*

In the case that there is no relationship between aggregate market news volume and abnormal returns, we hypothesize that this is due to the fact that on an aggregate level, markets are not affected by information processing constraints. In other words, limited attention is not impacting the aggregate market, and the market disseminates information efficiently.

In terms of firm-specific news volume, several news items can attract more investor attention that results in a faster dissemination of information, and therefore no long-term abnormal returns.[96] In other words, the relationship is the opposite compared to aggregate market news volume. Our hypothesis for firm news volume states as follows:

> ***H2c:*** *Firm news volume is negatively related to abnormal returns.*

In the case that there is a clear positive relationship between firm specific news and abnormal returns, we hypothesize that this is due to the fact that noise traders are attracted towards attention grabbing stocks, and their trades cause prices to deviate from fundamentals resulting in abnormal returns. In the case that there is no reaction, markets are efficient and firm specific news volume bears no correlation to returns.

3.2.3 Abnormal volume hypotheses

In terms of the link between trading volume and investor sentiment, there may be two scenarios. Either investors react more to the tone of the text, or alternatively to the information content within the text. Based on prior literature, we hypothesize that the latter effect dominates. Therefore the hypothesis:

> ***H3a:*** *Investor sentiment is related to a small increase, or no reaction, in abnormal trading volume over short time-periods, and with no-reaction or positive reaction in the long-term*

However, tone of text could also dominate over content. In such instance, we would expect sentiment changes to still have a very similar relationship to abnormal volumes as suggested in primary hypothesis. Investors would first trade based on tone, and then trade again to reverse their prior trades as the true nature of fundamental information is revealed.

Linking market news volume to trading volume is another interesting research question. We hypothesize that aggregate market news volume distracts investors, resulting in underreaction to information. We expect that this leads to a market where:

> ***H3b:*** *Market news volume has a negative relationship, or no relationship, with abnormal trading volume in the succeeding day, and a positive relationship with abnormal trading volume on the longer event windows.*

[96] Also, multiple ‘experts’ effect can be argued to take place when more news are present. The information credibility increased; therefore, decreasing the discounting of qualitative texts, and subsequently increasing the dissemination of information.

Alternatively, in the case of no relationship at all, the aggregate market may not be distracted by the amount of news flooding the market. In this case, the aggregate market would disseminate information efficiently, and not be affected by limited attention.

Next, we aim to link the firm specific news volume variable to trading volume. For this, we hypothesize that firm specific coverage increases attention towards a firm, resulting in faster information dissemination. This leads to the hypothesis:

> ***H3c:*** *Firm specific news volume has a positive initial relationship with abnormal volume followed by a declining, or neutral, relationship with the longer event windows.*

Alternatively, attention grabbing stocks could have a positive relationship with abnormal volume across all event windows as investors are more prone to buy attention stocks. Prior empirical evidence on the topic is mixed and we recognize that such hypothesis could be a viable alternative.

3.2.4 Abnormal volatility hypotheses

Researchers have also been interested in the link between volatility and sentiment. We expect sentiment changes to increase future idiosyncratic volatility according to the findings of previous literature. In line with prior literature, we hypothesize that noise traders exist in the market, and react to sentiment changes with exaggerated action:

> ***H4a:*** *Sentiment changes lead to increased abnormal volatility.*

In the case that sentiment changes do not explain idiosyncratic volatility variations, an alternative explanation could lie in our sentiment estimation methodology. In other words, our sentiment estimate would not be accurate enough, and would therefore introduce too large measurement errors that mask the true relationship.

In terms of linking volatility with market news volume, literature has not documented a clear relationship between market news volume and abnormal volatility. Therefore, we hypothesize that:

> ***H4b:*** *There is no significant relationship between market news volume and firm-specific abnormal volatility.*

In the case that we find a pattern linking aggregate market news volume to firm specific idiosyncratic volatility, we offer a preliminary hypothesis that such a relationship might proxy for market wide volatility that impacts firm specific volatility.

Finally, we are interested in firm specific news volume and volatility. For this, we expect that attention grabbing stocks attract more noise traders to trade on the stock:

> ***H4c:*** *Firm specific news volume increases abnormal volatility.*

In the case that there is no significant relationship between firm specific news volume and idiosyncratic volatility, we hypothesize that this is due to the fact that the market is sufficiently efficient to correct the trades of noise traders in a manner that attenuates idiosyncratic volatility. Therefore, no significant increases in idiosyncratic volatility would be detected.

In total, we proceed with ten hypotheses: one regarding the performance of our sentiment estimate, and nine regarding the link of media variables and stock performance. To summarize our expectations, we have gathered our hypotheses in the table below.

Table 1: Hypotheses of the study

Hypotheses from previous subsection are collected and briefly summarized in this table.

Sentiment estimation	
H1	Linearized Phrase-Structure -model outperforms the existing word-count based methodologies used for sentiment estimation in financial context.
Abnormal returns	
H2a	Investor sentiment forecasts future abnormal returns in all event windows through underreaction to sentiment changes.
H2b	Aggregate market news volume has a positive relationship with abnormal returns in longer event windows, while having a neutral, or negative, relationship with the 1-day event window.
H2c	Firm news volume is negatively related to abnormal returns.
Abnormal volume	
H3a	Investor sentiment is related to a small increase, or no reaction, in abnormal trading volume over short-term, and with no reaction, or positive reaction, in long-term
H3b	Market news volume has a negative relationship, or no relationship, with abnormal trading volume in the succeeding day, and a positive relationship with abnormal trading volume on the longer event windows.
H3c	Firm specific news volume has a positive initial relationship with abnormal volume followed by a declining, or neutral, relationship with the longer event windows.
Abnormal volatility	
H4a	Sentiment changes lead to increased abnormal volatility.
H4b	There is no significant relationship between market news volume and firm-specific abnormal volatility.
H4c	Firm specific news volume increases abnormal volatility.

4 DATA

In this section we describe our data. We start by describing our time period and the firm sample for which we have downloaded our variables. Next, we explain how we have chosen our main and control variables, and how we define them. Finally, we describe the process of collecting financial and media data for our sample as well as the sources we have used.

4.1 Time period and selected sample

We have selected to download our media sample between January 1st 2006 and March 31st 2011. We have chosen the period as it covers a full business cycle, including both sides of the crisis period of 2007-2009. As a result, our sample should resemble a full economic cycle, and therefore be unbiased from cycle dependent patterns.

During our sample period, the largest significant event in magnitude has been the financial crisis: starting with a housing boom fueled by innovative financial instruments, the credit market spiraled out of control and finally burst. Consequently, the stock market experienced a tremendous decline followed by several infamous bankruptcies from well-established and influential companies such as: Bear Stearns and Lehman Brothers. Especially the financial sector was shaken by the aforementioned collapses. Furthermore, as the credit crunch escalated, the financial crisis quickly developed into a full-blown economic crisis that struck a strong backlash to the entire real economy of the United States. For instance, car manufacturers were hit hard by the bankruptcies of Chrysler and General Motors. As a result, volatility and risk premiums hit record levels in late 2008 and early 2009. Moreover, with the intertwined banking systems, the U.S. crisis quickly spread across borders resulting in a global recession. The aftermath of the financial crisis continues even at the time of the writing when Europe struggles against the lack of confidence the markets has towards the Euro currency, and the sovereign government debt issues of European countries. The aforementioned has been dubbed as the Euro crisis.

In addition to the financial crisis, other significant global events include: the collapse of Iceland's economy, the anti-China protectionist movement in the U.S. fueled by China’s artificially low renminbi levels, highly volatile commodity markets and the largest Ponzi-scheme in the history: the Madoff scandal. Also, in terms of natural calamities: Iceland's volcanoes have restricted European flight traffic heavily, and Japan's earthquake and tsunami in March 2011 have also shaken the markets. At the end of our sample period, the Mideast uprising has been another dominating theme in the top news. The chart below illustrates how the stock market has performed during our selected time period. As can be seen from the chart, the time horizon covers fairly well a full economic cycle.

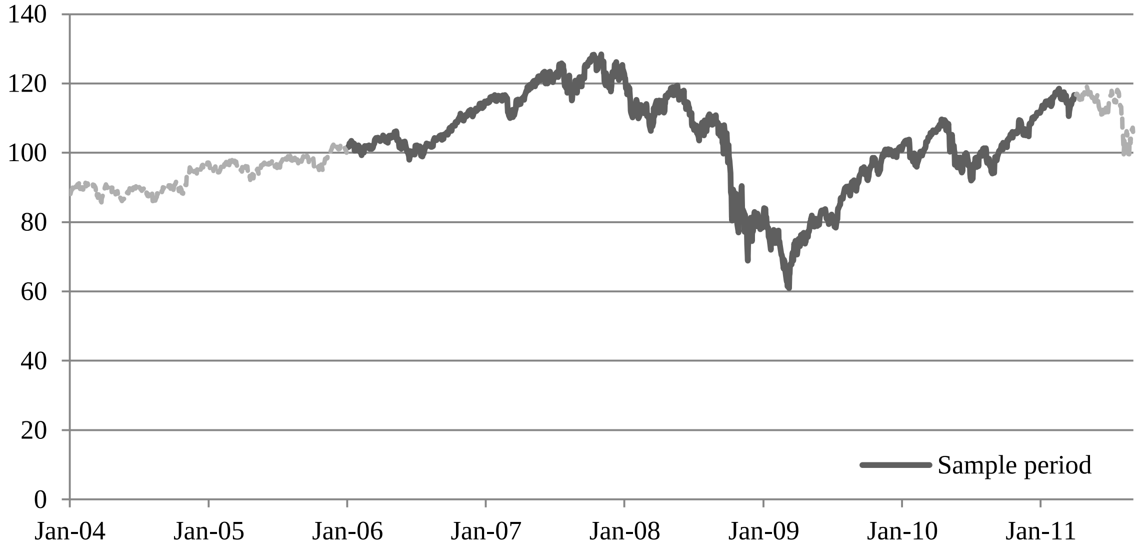


Figure 3: Development of S&P100 during sample period. Index=100 on 1 Jan 2006

For our sample, we have selected the S&P100 firms for their large coverage in news due to their strong status in the economy. Furthermore, with S&P 100 firms, the data on the stocks is easily obtainable, accurate, and the majority of the news are in English. The aforementioned is important as the majority of natural language processing technologies have been developed for English, and doing a study for another language would significantly complicate the detection of sentiment. Also, English is a fairly straightforward language for the purpose of analysis whereas other languages, such as Finnish, can be remarkably more burdensome to analyze due to postpositions in language. In Finnish, for instance: ‘*a language*’ translates as *‘kieli’*, while the expression *‘in a language’* translates into *‘kielessä’*, making the simple task of a word count much harder in Finnish vis-à-vis English. With English, the aforesaid challenge can be avoided.

We downloaded the list of S&P100 constituents and their tickers from Standard and Poor’s website on the 15th June 2011. All S&P100 constituents are large cap (market capitalization over one billion euros). The largest companies in the list, by market cap, on the 31st March 2011 are: Exxon Mobil, Apple and Chevron. The industry with the most companies is transportation.

Table 2: S&P100 summary statistics

Industry	# of companies
Transportation	10
Retail	8
Banking	8
Petroleum and Gas	8
Pramaceutcal Products	7
Computers	6
Telecommunications	5
Business Services	5
Utilities	5
Insurance	4
Other (17 industries)	34

Market capitalization, USDbn	
Average	79
Median	51
Maximum	414
Minimum	12

4.2 Study variables

In order to study the effects of investor sentiment on financial metrics, we need to establish what financial metrics we wish to study as well as how we will estimate investor sentiment.[97] Furthermore, we need to establish a proper set of control variables for our study so that our main study variables are not simply acting as a proxy for an omitted variable.

We will start by specifying the main variables of interest to our study: the dependent variables that capture a specific financial metric of interest and the key independent variables that we believe will explain variations in the dependent variables. After defining the main variables, we will move on to describe the set of control variables that we will include in our study to take into account the variables that are documented to have a relationship with our dependent variables. Finally, we also describe a set of additional control variables that we use later on for robustness checks.

For a summary of variables, we refer the reader to Appendix B – Main variable definitions, Appendix C - Main specification control variable definitions, and to Appendix D – Alternative specification control variable definitions. The aforementioned appendices define briefly our variables, and give some relevant prior literature reference examples.

[97] In Section 5.1 Estimating investor sentiment, we will discuss our definition of investor sentiment in more detail.

4.2.1 Main Variables

Previous literature on content analysis in the domain of finance and accounting has mainly focused on analyzing the relationship between an estimate of investor sentiment[98] and a specific financial metric. Five financial metrics that have been used most commonly include: raw returns[99], abnormal returns[100], trading volume[101], return volatility[102]and earnings[103]. The most common estimates for investor sentiment have included: a market news volume count, a firm specific volume count and a sentiment score, developed from news article through a content analysis method.

We have chosen to focus on the following financial metrics in our study: abnormal returns, trading volume and abnormal return volatility. Our focus is motivated by our interest in stock related metrics. In explanatory variables, we have chosen to include three different variables to capture — and dissect —the effect of investor sentiment on financial metrics. The variables chosen are: market news volume, firm news volume and an estimate for sentiment. We will describe our definition of each of the variables in more detail below. For discussion on our specifications, we refer the reader to Section 5.2.1., and for a summary of dependent and main variable definitions we refer the reader to section 4.2.

Abnormal returns

In order for us to estimate abnormal returns for our event windows, we need to define a benchmark return. Barber and Lyon (1997), concurrently with Daniel and Titman (1997), have shown that using matching portfolios based on size and book-to-market have produced more accurate test statistics than factor based models (i.e., Fama and French three-factor model: Fama and French, 1993). Recent literature on content analysis in the domain of finance has often relied on matching portfolio methodology for calculating abnormal returns (e.g. Chan, 2003; Engelberg, 2008; Hirshleifer et al., 2009; Demers and Vega, 2010). Hence, we adopt Fama and French's (1992) matching portfolio approach to calculate benchmark returns. The method divides the market into a number of portfolios according to company size (market equity) and book-to-market ratio. The idea is that companies with the same size and

[98] Also referred to as linguistic 'tone' or 'style' (e.g., Davis et a., 2008; Loughran and McDonald, 2011)
[99] E.g., Chan, 2003; Antweiler and Frank, 2004; Tetlock, 2007; Tetlock et al., 2008
[100]E.g., Chan, 2003; Li, 2006; Engelberg, 2008; Tetlock et al., 2008; Hirsleifer et al., 2009; Demers and Vega, 2010; Loughran and McDonald, 2011
[101] E.g., Antweiler and Frank, 2004, 2006; Tetlock, 2007; Hirsleifer et al., 2009; Loughran and McDonald, 2011
[102] E.g., Antweiler and Frank, 2004; Demers and Vega, 2010; Loughran and McDonald, 2011
[103] E.g., Li, 2006, 2008; Tetlock et al., 2008; Loughran and McDonald, 2011

book-to-market ratios should yield same returns. As a robust check, we replicate our results using alternative definitions[104] for returns.[105]

To calculate the benchmark returns for the matching portfolios, we use daily returns that are retrieved from Kenneth French's website.[106] The 25 matching portfolios, which have been constructed at the end of each June, are the intersections of 5 portfolios based on size: market equity [ME], and 5 portfolios formed on the ratio of book equity [BE] to market equity [BE/ME]. The size breakpoints for a given year are the NYSE market equity quintiles at the end of June of the respective year. BE/ME -ratio for June of a given year is the book equity for the last fiscal year end before the respective year divided by the market equity for December of the preceding year. The BE/ME breakpoints are the NYSE quintiles. As a weighting scheme, we choose to use equal weights in line with previous literature that has disclosed the used weighting scheme (e.g., Chan, 2003; Demers and Vega, 2010).[107]

To match our firms with a correct portfolio, we use the market equity [ME] and book-to-market values [BE/ME] for each S&P100 companies. Hence, we can find the corresponding portfolio for each of the S&P100 companies. As we are dealing with S&P100 firms, 99.5% of our observations are ranked within the largest category in terms of market equity, and all the observations fit within the three top categories of market equity.

To calculate event period abnormal returns, we use a buy-and-hold approach in an attempt to better capture the true impact for an investor vis-à-vis using cumulative abnormal returns [CARs], or other variants of return methodologies. Buy-and-hold approach has been used in recent literature by Loughran (2011) and Hirsleifer et al. (2009)[108].

When calculating abnormal returns, we use close-to-close prices in our event window. Therefore, event window [0,1] would read as: share bought at the closing price of day 0 and

[104] For more discussion on alternative methodologies for returns, see Section 5.4.2.

[105] We test raw returns (e.g., Tetlock, 2007) and alternative abnormal return benchmarks: value weighted S&P 100 returns (e.g., Loughran and McDonald, 2011) and the Fama and French three-factor model (e.g., Tetlock et al,. 2008). Our results are qualitatively similar with marginal changes in coefficients' statistical significance.

[106] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html - retrieved on the 25th July 2012.

[107] Our results are robust for using value weight returns, and do not disappear with the change of weighting scheme as suggested by Fama (1998). However, the magnitude of our results diminishes to some extent when using value weights.

[108] To take into account the critique directed towards results from buy-and-hold strategies (e.g.,; Mitchell and Stafford, 1997; Fama, 1998), we replicate our results using cumulative abnormal returns [CARs] as in Engelberg (2008) and Demers and Vega (2010). In line with the critique of Fama (1998) and Mitchell and Stafford (1997), we find that our coefficient estimates decrease in magnitude when using CARs instead of buy-and-hold returns. However, our results remain qualitatively unchanged.

sold at the closing price of day 1[109]. To study the relationship between abnormal returns and our main variables, we define several different event windows: first, we will study the abnormal performance of a firm on the following day after the event day [0,1] (e.g., Tetlock, 2007; Tetlock et al., 2008; Hirsleifer et al., 2009); second, we will expand our event window slightly to 4 days [1,5] to examine whether or not a prolonged short-term effect is in play (e.g., Griffin, 2003; Engelberg, 2008; Demers and Vega, 2010; Loughran and McDonald, 2011); third, we will move on to analyze intermediate effect by modifying our event window into a 30-day window [2,32]; finally, we will examine the potential long-term impact by broadening our event window into 60 days [2,62] (e.g., Engelberg, 2008; Hirsleifer et al, 2009; Demers and Vega, 2010).

All of our windows exclude the event date returns as we cannot be certain when the event occurs during the day. For example, let us illustrate our point through trading strategy example: if we know that the sentiment for the event day [0] is significantly negative, we would want to short the stock in question. However, in the absence of accurate time stamps on news, and intraday stock price data, we cannot open a position based on the opening price of the event date, as that would imply that we would “see into the future”. We do not know when the sentiment turned negative during the day or after market close: all we know that the sentiment for the day was negative. Hence, for example, the sentiment might have turned negative before midnight with the release of few influential articles in Asia. As a result, we begin all of our event windows after day 0 in all windows to be certain that we are comparing returns that take place after the sentiment change, and the causality of events is correct.[110]

Previous literature has used a multitude of event windows to study the potential relationship with sentiment and returns. All of the above example references do not— per se— use the same exact event windows. However, they all study a similar time horizon: short, intermediate or long. In order to avoid data dredging concerns expressed by the proponents of

[109] When comparing results with different studies, the reader should be aware that some studies report event windows using opening and closing prices. Therefore [0,1] would read as: share bought at the opening price of day 0 and sold at the closing price of day 1 that would equal - in our definition - an event window of [-1,1].

[110] As we do not have access to intraday data like Tetlock (2007) and Tetlock et al. (2008), we cannot judge market efficiency with the short-term window of [0,1] as it is possible that markets have reacted instantaneously to new information. Furthermore, for majority of the news, LexisNexis stores only the date of the news. Thus, we do not have information on the actual release time of the news or even the time zone of the date. This makes it impossible to study precisely the immediate stock reaction to the release of a news item. In other words, we cannot separate when the actual sentiment change has taken place during day 0. Therefore, sentiment change might have taken place after the closing price of day 0, and the new information would then be immediately reflected in the opening price of day 1. As a result, event window [0,1] is aimed at establishing a link between sentiment and financial metrics, not to draw inference on the speed of information dissemination.

EMH with subjective event windows (e.g., Fama, 1991), we employ several different event windows that have been designed to be as objective as possible. As discussed, we employ event windows [0,1], [1,5], [2,32] and [2,62] for abnormal returns.

Abnormal volume

In line with the recent literature in content analysis (e.g. Tetlock, 2007; Hirsleifer et al., 2009; Loughran and McDonald, 2011) we are examining the effects of investor sentiment on trading volume. Consistent with the previous research, we define trading volume in terms of abnormal trading volume. In the absence of a consistent definition in previous literature for measuring abnormal trading volume, we rely on the definition used by the most recent publication of interest: Loughran and McDonald (2011). Therefore, we define abnormal trading volume for an event as the sum of the daily abnormal volumes for the event window. In other words, for event window [2,5] the abnormal volume would be the sum of the event day abnormal volumes during the time period (sum of abnormal volumes for days [2],[3],[4] and [5]. Abnormal volume for a day is defined as the difference of the event day trading volume and the mean trading volume for [-62, -2] divided by the standard deviation of trading volumes for [-62, -2]. To summarize, we calculate abnormal volume for day t as:

$$\text{Standardized abnormal volume for day}_t = \frac{V_t - \mu_V}{\sigma_V} = \frac{V_t - \frac{\sum V_{-62\ldots-2}}{60}}{\sigma(V_{-62\ldots-2})}$$

where V_n represents the absolute trading volume at time n.

As we do not foresee any specific reason as to why we should employ different event windows with volume vis-à-vis returns, we run our specification with the same four event windows discussed in the previous section in order to maintain consistency within the study. We calculate abnormal volume for the same event windows as for returns: [1], [2,5], [3,32], [3,62]. Note that the notation changes slightly with volume vis-à-vis returns even though the event window remains the same. The reason is that with returns we are using closing prices and with volume we are reporting the actual days that volume is counted in the event window. Therefore, [0,1] in returns would report a return for 1-day period utilizing closing prices whereas volume notation [1] reports the same thing: volume for 1-day period.

Abnormal volatility

Following the studies of Demers and Vega (2010) and Loughran and McDonald (2011), we study the impact of investor sentiment on the subsequent volatility of abnormal returns.[111] As we are not interested in systematic volatility, we use abnormal returns to isolate the idiosyncratic volatility related to a given firm. We define the volatility of abnormal returns as the standard deviation of daily abnormal returns for the event window in question.

Demers and Vega (2010) define abnormal return volatility as the logarithm of the sum of squared abnormal daily returns during the event window while Loughran and McDonald (2011) define abnormal return volatility as the root-mean square error from a Fama-French (1993) three-factor model. Our results are not sensitive to changes in the definition of abnormal return volatility.

To assess abnormal idiosyncratic volatility, we rely on two of the same event windows we have used before: 30-day window [2,32] and 60 day window [2,62]. The shorter event windows have been excluded for the lack of observations in which to base the volatility calculations.[112]

Market news volume

Hirsleifer et al. (2009) find that simultaneous earnings press releases distract investors, resulting in stronger post-announcement drifts on days when there are large quantities of news released vis-à-vis days when there are low quantities of news released. Therefore, we include a market news volume variable to capture the distraction effect demonstrated by Hirsleifer et al. (2009).

We define the market news volume as the number of all the different media items for the given firms for a given day. To elaborate further, if there are three firms in the market [A, B, and C], and firm A has 3 different news articles on day 1, and firms B and C have 1 news items each on day 1, then the market news volume would equal 5 for day 1 [= 3+1+1].

Also, we run an alternative specification with standardized market news volume to see if seasonality is driving our results with the variable. Furthermore, we test for the impact of

[111] Abnormal returns are defined against the benchmark of 25 size and book-to-market matched portfolios

[112]We believe that the 4-day event period would provide too much noise in the estimate of volatility to provide any useful insight into the relationship with our main variables. Our approach differs in this stance from that used by Demers and Vega (2010) who employ a short-term event window in addition to the long-term window.

abnormal market news volume using previous year's average volume. We define the standardized form of the variable for day 0 as:

$$\frac{\left(News\ Volume_{0} - \ Average\ Volume_{-252,-2}\right)}{Standard\ Deviation\ of\ Volume_{-252,-2}}$$

Firm news volume

Some recent studies have suggested that firm-specific media coverage can affect a firm's returns (e.g., Fang and Peress, 2009; Loukusa, P, 2011). Hypotheses for the suggested link vary. For instance, Demers and Vega (2010) suggest that the presence of multiple experts: different sources of same information, can affect the impact that investor sentiment has on abnormal returns. Therefore, the presence of multiple news articles on a given day could impact abnormal returns. Also, firm specific news coverage could potentially represent the attention directed towards a company. Therefore, in line with limited attention, it could signal that the company has more attention directed towards it and therefore experiences less underreaction. In spite of mixed evidence, we include a proxy for firm-specific media coverage to study the potential effect. We define the firm news volume for a given day as the number of the different news items for a given firm on a given day. In example, if a firm has 3 different news items on a given day, the firm news volume variable would take the value of 3 for that day.

As with market news volume, we run an alternative specification[113] with standardized firm specific news volume to test the impact of abnormal news volume with seasonal adjustment. We define the standardized form of the variable for day 0 as:

$$\frac{\left(News\ Volume_{0} - \ Average\ Volume_{-252,-2}\right)}{Standard\ Deviation\ of\ Volume_{-252,-2}}$$

Sentiment estimate

We discuss our primary sentiment estimation methodology in more detail in: Section 5.1. However, here we will briefly discuss the two different approaches we use for estimating investor sentiment: '*bag-of-words'* vector words counts, and the idea of Linearized Phrase-

[113] More on this in Section 6.3.2

Structure -model[114]. Linearized Phrase-Structure -model is our novel method for estimating investor sentiment that will be the primary method used in our study.

To capture the sentiment of a news article, several studies have attempted to use various content analysis methods with the most common used method being a word count (i.e., '*bag-of-words*' vector word count) based on a predefined dictionary.[115] Word count method relies on an intuitive assumption that a positive text includes more positive words, and a negative text includes more negative ones. Thus, by counting the positive and negative words in a news article, an article with many positive words should be positive, and vice versa.

Counting words is a relatively straight forward task: an algorithm is given a list of words from a dictionary, and it calculates the frequency of each word in an article. To ensure that we count all the correct occurrences, we either need to take into account all inflection forms (e.g. *'bull'* and *'bullish'*), or alternatively convert all words to their lemma form (e.g., *'bull'*). Furthermore, in the case of inflections, it is critical to make sure that only exact matches are compared against the list to avoid double counting. For instance, searching with *'bull'* could find both *'bull'* and *'bullish'*, and *'bullish'* could find *'bullish'* again. Therefore, double counting would occur in this instance.

The most established used dictionary so far has been the Harvard Psychology Dictionary (H4N).[116] In fact, since the seminal study of Tetlock (2007), in which Tetlock shows that negative word counts using the Harvard Psychology negative word category divided by the total number of words for a news article explain most of the variation in returns, the majority of studies has focused on the fraction of negative words as defined by the H4N.[117]

However, the caveat with using word lists is the domain-specificity of dictionaries: the meaning of a word is domain specific in many instances. In fact, Loughran and McDonald (2011) show that in financial context the H4N dictionary misclassifies approximately 73% of

[114] When testing for efficiency of our categorization algorithm, we benchmark this also against quasi-compositional sequencing. As this method has not been used in previous finance literature and is clearly outperformed by the LPS, we do not use the method when testing for sentiment and stock performance.

[115] e.g., Li, 2006; Tetlock, 2007; Bligh and Hess, 2007; Engelberg, 2008; Davis et al., 2008; Tetlock et al., 2008; Demers and Vega, 2010; Loughran and McDonald, 2011.

[116] e.g., Tetlock, 2007; Engelberg, 2008; Tetlock et al., 2008; Demers and Vega, 2010; Loughran and McDonald, 2011.

[117] Tetlock (2007) constructs a 77x77 covariance matrix from all the categories in the Harvard dictionary in order to construct a linear combination variable that captures most of the variation in the matrix based on the 77 Harvard dictionary categories. However, after constructing the variable, and studying its impact, Tetlock concludes that simply using the negative word category is a fair proxy for sentiment that captures the variation in an equivalent manner vis-à-vis the linear combination variable.

the words as being negative when in fact they are not. For example, *'tax'* is regarded as a negative expression in Harvard dictionary. However, when reading a financial statement, the word *'tax'* does not have a negative meaning but simply describes a firm's tax position. In fact, Loughran and McDonald point out that misclassifications can induce spurious correlations, and therefore transcend the status of simply adding noise to the estimate. Therefore, researchers should use domain specific dictionaries when employing word counts. For that purpose, Loughran and McDonald develop a dictionary with six different word lists to be used in financial context.

The idea that dictionaries are crucial for the success of content analysis is not a novel one. In fact, to paraphrase Berelson (1952): content analysis stands or falls by its categories. Hence, we will test both of the aforementioned dictionaries in our study in an attempt to evaluate the efficacy of the dictionaries as well as to compare them with our primary methodology.[118] For a description of how we aggregate sentiment of multiple news items on a day, we refer the reader to Section 5.2.

Our primary method for estimating investor sentiment: Linearized Phrase-Structure -model, is a novel method developed for this study. We employ a natural language programming algorithm that captures sentiment on a sentence level, and then analyzes the sentiment of the article by looking at the aggregation of sentence sentiments inside the article. We refer the reader to Section 5.1 for more information on investor sentiment estimation with the various methods.

4.2.2 Control Variables

In Section 2.1.3, we described several known anomalies that have been shown to explain abnormal returns in contradiction to the efficient market hypothesis [EMH]. In order for us to study the effect of investor sentiment on financial metrics, we must control for the impact that other known variables have with our dependent variables. Therefore, we include a set of control variables into our study.

[118] We will use the negative fraction of words for the sake of consistency with prior studies even though Loughran and McDonald (2011) compellingly argue for the use of term weighting functions. Furthermore, we refrain from standardizing the negative fraction as we believe that our sample might exhibit systematic increase in negativity due to the financial crisis that should not be smoothened out. Furthermore, Engelberg (2008) finds similar results with and without standardization; therefore, we do not foresee a problem with our choice of methodology.

Our underlying goal for selecting controls is to be consistent with previous literature and to create a comprehensive set of controls in order to mitigate the risk of an omitted variable driving our results. Our controls include both dummy (binomial) and continuous number variables.

We begin by discussing controls used in our main specifications. From there on, we move on to describe the additional controls we employ in our alternative specifications as robustness checks. Once again, we refer the reader to Section 5.2.1. for more discussion on our study specifications.

Main specification controls

Most of our control variables are designed to control for the different documented anomalies relating to abnormal returns.[119] However, prior literature has commonly used the same controls with other dependent variables as well. Therefore, we employ a similar set of controls to all our main specifications with the exception of abnormal volume specification where we add abnormal market trading volume to the set of controls. More information on the main specifications of our study can be found in Section 5.4.1., and for a summary of the variables discussed in this section we refer the reader to Appendix C - Main specification control variable definitions.

Size

In order to control for the size effect discovered in previous studies on cross-sections of returns, we include a control variable for firm size. We define firm size as the log of market equity. Market equity is taken as reported on Datastream. Our definition of size follows that of Tetlock et al. (2008), Hirsleifer et al. (2009) and Loughran and McDonald (2011), among others. By controlling for size we follow the mainstream approach of modern finance research.

Book-to-market

In line with modern finance research and the most recent influential content analysis papers, we control for book-to-market in our specifications. We define book-to-market as the log of:

[119] In truth, as we are using matching portfolios as benchmark returns, several of our control variables should be implicitly controlled by the benchmark returns. Nevertheless, we include such controls in order to be as robust as possible.

one plus market equity divided by the last reported book value of equity. This is in line with, for instance: Tetlock et al. (2008) and Hirsleifer et al. (2009).

Momentum

Since the seminal articles by DeBondt and Thaler (1985) and Jegadeesh and Titman (1993), research has documented a relationship between past returns and future abnormal returns.[120] The two different effects have been named reversal effect and momentum effect. To investigate the impact of past returns on future abnormal returns we include three different controls of past abnormal returns:

- Short-term abnormal returns [-4,-1]
- Intermediate-term abnormal returns [-34, -4]
- Long-term abnormal returns [-255, -34]

Our event windows for the momentum variables are similar to those of Tetlock et al. (2008). Other influential content analysis papers have mainly relied on one momentum variable that captures the past returns of a year (e.g., Loughran and McDonald, 2011) but we choose to dissect the one year variable into several different variables to study the impact of past returns in more detail. Abnormal returns are defined as described in Section 4.2.1.

Share turnover

In line with the previous mainstream finance literature, and recent content analysis papers (e.g., Tetlock et al., 2008; Hirsleifer et al., 2009; Demers and Vega, 2010; Loughran and McDonald, 2011), we include share turnover to our controls in order to control for the liquidity of the stock and belief dispersion concerning the stock (e.g., Hong and Stein, 2003), and the subsequent impact on returns. We define share turnover as the log of: 1 + sum of the volumes for [-252, -2], divided by shares outstanding at event date. Our definition is similar to that used by Loughran and McDonald (2011).

Standard unexpected earnings [SUE]

In order to control for the well-known post-earnings announcement drift [PEAD], we include standard unexpected earnings variable. Previous literature on content analysis in finance domain has consistently included SUE in their controls, or as the dependent variable in their

[120] Refer to Section 2.1.3. for more information on the findings of prior literature.

specifications (e.g., Li, 2006; Tetlock et al., 2008; Davis et al,. 2008; Hirsleifer et al., 2009; Demers and Vega, 2010; Loughran and McDonald, 2011). However, previous research has defined SUE in myriad ways. The significant difference between definitions culminates in the scaling / standardization procedure used for SUE. Tetlock et al. (2008) and Demers and Vega (2010) use the time series method of Bernard and Thomas (1989) to standardize unexpected earnings with the standard deviation of the previous 20 quarters of unexpected earnings. Davis et al. (2008), Hirsleifer et al. (2009) and Loughran and McDonald (2011), on the other hand, standardize unexpected earnings using the stock price of the firm. We rely on the latter methodology as it has gained more exposure on recent influential studies, and is more pragmatic to implement.

We define SUE as the difference between the last reported quarter's EPS and the corresponding last median analyst forecast for that EPS divided by the closing share price on the day of the respective earnings announcement.

Abnormal volatility

If investor sentiment has predictive power over abnormal returns in contradiction with the prediction of the efficient market hypothesis, we must establish whether or not this is due to limits to arbitrage.[121] As pointed out by Shleifer and Vishny (1997), agents may not trade on information they have if they face constraints.[122] In order to evaluate whether agents face constraints, we include a control for idiosyncratic volatility: the most common proxy for arbitrage risk used in main stream finance literature (e.g., Engelberg, 2008). We define abnormal volatility as the standard deviation of daily abnormal returns for the time period of [-252, -2].

Institutional ownership

Some researchers (e.g., Frazzini, 2006) have suggested that irrational trading by institutions is a source of post-earnings announcement drift [PEAD]. Hence, institutional ownership could be a determinant of abnormal returns. Also, according to the limited attention theory[123], institutional ownership can have an impact on abnormal returns. As investors' attention is limited, they face information processing costs. Therefore, investors with more processing

[121] See Section 2.1.4. Limits to Arbitrage for more information on the topic.

[122] Also, prior literature has shown that uncertain firms: measured by volatility or the nature of their environment, face more severe underreaction, and hence are impacted more heavily by investor sentiment (e.g., Brav and Heaton, 2002; Dye and Sridhar, 2004; Engelberg, 2008)

[123] See Section 2.2.1. for more details.

capacity should be ahead of the curve and capture abnormal returns. Previous literature (e.g., Engelberg, 2008) assumes that institutions have more processing capacity than individuals. Therefore, in light of limited attention theory, institutions should earn abnormal returns. Moreover, in light of our study, investor sentiment should have less predictive power over financial metrics with firms that have high institutional ownership. We define institutional ownership as shares held by government, pension funds and investment companies divided by the total shares outstanding[124].

Abnormal market volume

In order to study the effect of investor sentiment on abnormal trading volume, we must isolate systematic jumps in trading volume. In line with Antweiler and Frank (2006) and Hirsleifer et al. (2009), we include a control for abnormal market volume in the abnormal trading volume specification. We define abnormal market volume in the same way we defined abnormal volume in Section 4.2.1: the sum of daily abnormal volumes for a given event period. Daily abnormal volume is defined as the difference of daily trading volume and mean volume, divided by the standard deviation of volume. Mean volume and standard deviation of volume are defined based on days [-62,-2]. The aforementioned methodology is consistent with Loughran and McDonald's (2011) methodology for calculating abnormal trading volume.

To establish a proxy for market trading volume, we take the sum of trading volumes of all our firms in a given data set (i.e., S&P 100 firms). From there on, we employ the aforementioned definition of abnormal trading volume to calculate abnormal market trading volume.

Alternative specification controls

In order to make sure that our results are robust, we include several additional controls that have been used in some of the studies of prior literature. However, most of the additional controls have had little impact in prior studies, and for the most part have been insignificant. Therefore, we do not expect to see a significant impact from the additional controls. Nevertheless, for the sake of robustness, we include the additional controls in our alternative specifications to study their potential impact on our results. Discussion on our alternative specifications can be found in Section 5.4.2., and for a summary of the variables in this section we refer the reader to Appendix D – Alternative specification control variable definitions.

[124] The definition excludes hedge funds, as the data is not available to us from the databases we have access to.

Industry dummies

To control for the industry specific risk factors we employ the 48 different industry classifications suggested by Fama and French (1997) in their seminal article. Recent content analysis papers in the domain of finance have also employed similar methodology to classify firms into different industry segments (e.g., Engelberg, 2008; Loughran and McDonald, 2011). Using the 48 distinct Fama and French categories, we classify our firms into different industries based on their primary industry classification codes (SIC), as retrieved from Datastream. After categorizing our firms, our sample is divided into 28 different Fama and French industry categories (S&P100 does not have firms for the 20 remaining industries).

# of analysts following

Analyst coverage is often used as a proxy for media coverage and informational efficiency (e.g., Demers and Vega, 2010). The underlying notion is that firms with higher levels of analyst following have simply more information available, and hence exhibit lower abnormal returns.

In line with Engelberg (2008), Hirsleifer et al. (2009) and Demers and Vega (2010), we include a control for analyst coverage. We define the # of analyst following in line with the aforementioned authors as the log of: the sum of 1 and the number of analysts following a given firm.

Analyst dispersion

We include analyst dispersion to control for the prevalent belief dispersion that can impact financial metrics according to previous evidence. Recent content analysis research has also included analyst dispersion in controls (e.g., Tetlock et al., 2008; Demers and Vega, 2010; Loughran and McDonald, 2011). We define analyst dispersion as the standard deviation in analysts’ prior EPS estimates for the most recent reported EPS, divided by the share price on the respective earnings announcement date. This maintains consistency with SUE estimate, and our definition is similar to that of, for example: Loughran and McDonald’s (2011) definition.

Calendar dummies

To control for the well-documented January, Monday and end-of-the-month effects, we include control variables for the aforementioned. In the context of calendar dummies, the dummy will take the value of 1 if the variable in question is located within the event period. For instance, in the case that the day Monday is present in an event window; the dummy will assume a value of 1.[125] Calendar dummies have been used in recent content analysis papers in the domain of finance by Tetlock (2007), Li (2008) and Hirsleifer et al. (2009).

Last twelve months [LTM] dividends

In order to capture the documented dividend effect on a firm's performance, we include the announced dividends for the last twelve months [LTM] divided by the last reported book value of equity (e.g., Fama and French, 2006; Li, 2006).

No paid dividends during last twelve months [LTM] dummy

In order to capture the potentially significant constant impact of no dividends vis-à-vis small dividends, we include a control variable for firms that have paid no dividends during the last twelve months (e.g., Fama and French, 2006; Li, 2006). The variable will take the value of 1 if firms have paid no dividends during the last twelve months.

4.3 Financial dataset

For assessing stock performance, we retrieve stock prices and other necessary quantitative factors, and calculate all financial variables described in section 4.2. In this chapter, we explain how we retrieve the data and detail out calculation of these different metrics.

We retrieve data using Datastream Excel add-in that finds the data in question based on a company's ticker identifier. As the tickers Datastream uses are different from S&P 100 tickers, we manually match Datastream tickers with their corresponding S&P tickers. From there on, we download the data for the firms. Once downloaded, we remove all the data for non-trading days as non-trading days are not included in our analysis, as there are no relevant changes in financial metrics for non-trading days.

To make sure our price data obtained from Datastream is reliable, we conduct a number of sanity checks. To see large jumps in the data, we look at daily returns that are out of the range of -50% … +50% as such jumps are unexceptional. The method yields us two huge price changes with Citigroup and Morgan Stanley. Looking into the two extreme observations (see Appendix E – Outlier values in financial data) reveals that the jumps are in fact correct, and

[125] The inclusion of each calendar dummy is subject to the event window for the apparent reasons.

there is no error in data. We also check that market capitalizations and P/B variables are reliable: i.e., that the largest companies according to common knowledge have also the largest market caps, and that P/B figures are not extreme. Furthermore, we perform a check on market capitalization that is similar to the one we did with share price to see if there are errors in data. At first, share price and market cap checks may sound interchangeable. However, M&A and capital structure changes (changes in number of shares) create differences between the two. Therefore the list of changes (see Appendix E – Outlier values in financial data) is not identical to the changes in share price.

Based on our analysis of key changes in market capitalization, we correct for Lowe’s 3-week doubling of capitalization and Visa’s doubling of value one week after its listing. After the aforementioned corrections, we remain assured that there are no huge unexplained jumps in the data: either there has been a major event, or a more technical change has been adjusted by Datastream in the stock price. Table 3 shows descriptive statistics for our main specification variables for the entire sample.

Table 3: Main specification descriptive statistics[126]

	Average	Min	Median	Max	Interquartile range	Standard Deviation
Dependent Variables						
Abnormal returns						
[0,1]	0,01 %	-37,50 %	-0,04 %	75,78 %	1,56 %	1,91 %
[1,5]	0,02 %	-47,83 %	-0,07 %	98,62 %	3,24 %	3,73 %
[2,32]	0,06 %	-62,73 %	-0,15 %	215,09 %	9,46 %	9,38 %
[2,62]	0,07 %	-77,44 %	-0,41 %	229,18 %	13,56 %	13,19 %
Abnormal volume						
[1]	0,08	-4,11	-0,20	79,52	1,16	1,44
[2,5]	0,30	-10,91	-0,52	145,04	3,92	4,23
[3,32]	2,14	-38,02	-0,07	166,32	21,25	16,60
[3,62]	4,18	-64,99	2,64	165,01	29,10	22,55
Abnormal volatility						
[2,32]	1,58 %	0,32 %	1,27 %	17,50 %	0,91 %	1,08 %
[2,62]	1,60 %	0,44 %	1,30 %	13,26 %	0,88 %	1,03 %
Sentiment						
LPS*	17,54 %	0,00 %	16,69 %	100,00 %	17,79 %	13,48 %
Wordcounts						
Finance dictionary	1,17 %	0,00 %	1,02 %	15,09 %	0,98 %	0,87 %
H4N dictionary	2,41 %	0,00 %	2,38 %	19,92 %	1,15 %	1,00 %
Main Independent Variables						
Market news volume	359	108	306	1053	120	161
Firm news volume	4	0	1	235	4	7
Control Variables						
Size	4,70	3,48	4,66	5,72	0,50	0,36
Book-to-market	0,15	-2,00	0,14	1,41	0,11	0,10
Momentum						
[-4,-1]	0,02 %	-49,83 %	-0,06 %	93,21 %	2,78 %	3,26 %
[-34,-4]	0,07 %	-67,93 %	-0,22 %	163,51 %	9,40 %	9,41 %
[-255,-34]	0,21 %	-88,91 %	-2,02 %	367,30 %	24,45 %	24,57 %
Share turnover	0,50	0,00	0,47	1,43	0,22	0,19
SUE	0,00	-0,47	0,00	0,28	0,00	0,02
Abnormal volatility	1,67 %	0,55 %	1,41 %	8,47 %	0,93 %	0,93 %
Institutional ownership	6,96 %	0,00 %	6,00 %	64,00 %	11,00 %	7,96 %
Abnormal market volume						
[1]	0,06	-4,18	-0,06	6,00	1,41	1,22
[2,5]	15,42	1,54	14,83	38,88	6,51	5,43
[3,32]	149,88	72,92	145,88	356,43	59,54	44,56
[3,62]	306,78	153,70	303,51	704,46	129,75	89,30

* The maximum value of the LPS-model is 100% due to a few biased news items that include a large number of special characters and only one sentence. While we have removed short messages and messages with only tables, detecting patterns of various special characters would be significantly more diffcult. As the volume of such news items appears small, we leave these news into our sample.

[126] For variable definitions, refer to Section 4.2.

4.4 Media dataset

In this section we explain how we have collected our media sample that we use in the estimation of sentiment. We describe the sample we have chosen, and how we have downloaded the data. Also, we describe the preprocessing steps that we need to take before using the data texts in sentiment calculations. Finally, we summarize our news with descriptive statistics and figures.

4.4.1 Media sample selection

When collecting the media sample, one must first make a choice on what sources of qualitative texts to include. Using a single source of qualitative text as a sentiment proxy is the simplest way in terms of conducting the study. For instance, Tetlock (2007) uses only one source of qualitative text: the Wall Street Journal Abreast of market column, as a proxy for investor sentiment in his seminal study. However, aggregating multiple media items per day when estimating investor sentiment is the next step forward as investors do not rely on a single source of information when forming their opinions. For example, Tetlock et al., (2008) use multiple media sources in their study forcing them to pool different extracted sentiment scores from different media items to come up with one sentiment score for the day. Indeed, using multiple sources of qualitative text does reduce noise and improve accuracy vis-à-vis a naïve sentiment estimate from only one source (Das, 2010). Therefore, we also adopt this approach when collecting our media sample.

We aim to create as comprehensive media sample as is possible with our accessibility to data. To do so, we collect all the news feeds and earnings announcements accessible to us. Our media sample is not a comprehensive representation of all media, as it excludes such sources of sentiment as: twitter feeds and other forms of social media, TV-broadcasts, and internet message boards. Nevertheless, these sources have yielded weak results in prior literature (e.g., Antweiler and Frank, 2004; Das and Chen, 2006), or have been left unexplored to this date due to the large noise inherent in the sentiment data extracted from these sources.[127] To summarize, we create our sentiment from a feed of news and earnings announcements for three main reasons:

[127] Extracting accurate estimates of sentiment from social media is very demanding as the language used is less conventional, and much of the information conveyed in social media can have no direct link to a company's name but can instead deal with major products etc. For instance, people complaining about 'iPhone' but not mentioning 'Apple' in their messages.

- News and earnings announcements form the most significant portion of a sentiment. They are the key sources of information that an investor following a stock would study, and they represent the largest volume of up-to-date information on a company.
- News and earnings announcements can be retrieved from databases, covering thousands of sources. Collecting twitter feeds, TV-broadcasts, analyst reports etc., could add even further information to the sentiment, but these sources are difficult to retrieve. Furthermore, the critical information in the aforementioned sources is likely to be reflected in written news – at least after a lag.
- The format of public official written qualitative texts is an important advantage when analyzing texts. Public texts are often written in a format which avoids colloquial language. Hence, analysis of news is easier as we can use a standard established dictionary.

For these reasons, we choose to focus on news and earnings announcements. While it would be interesting to combine our sentiment to additional sources of different media, we leave this area for future researchers to explore.

4.4.2 S&P100 News and earnings announcements

Our dataset covers news and earnings announcements for all S&P100 companies. For these companies, we have gathered the news in major sources written during our sample period: between January 2006 and March 2011.

The news stories and earnings announcements for S&P100 companies were gathered from the LexisNexis database, with the source setting being '*Major World Publication'* [128] These sources include a comprehensive set of 624 different world publications from ABIX - Australasian Business Intelligence to Zimbabwe Standard (Harare). While it would be possible to extend the dataset even further with the selection of: '*All News (English)*' by 3 000 more news sources, the data set that we are retrieving with the Major World Publications is vast, and we consider it to be sufficiently representative of news publications that have an impact on investor sentiment.

LexisNexis allows searching for different functionalities, such as the options to look for keywords, to specify different time periods, select sources etc. A sample interface for LexisNexis search is included in Appendix F – Details on gathered media data, in section

[128] Major News Publications, as defined by LexisNexis: "This includes the world's major newspapers, magazines and trade publications which are relied upon for the accuracy and integrity of their reporting."

LexisNexis search interface. The easiest way for a user to download all relevant news would be to simply type company name in the search bar, select the relevant time period and sources, download the news, and repeat this for each of the companies. Unfortunately the web service does not work as easily as this, and we face two challenges: the difficulty in searching for news of a particular company, and limits on news download volume.

First, finding relevant news for a particular company is not a straight-forward task: searching for a company name results in a wide range of data. In example: searching for the company Apple with word '*apple*' yields the results that are related to the company. However, it also extracts unrelated news items on apples in juices etc. While some company names are rather unique, counter examples are numerous. The perfect solution would be to search for a company's known name in the news to include all relevant news, and afterwards sort out the irrelevant news. However, the approach results in excessive data that would need to be reclassified. The alternative to searching by names is provided by existing classifications in the databases: LexisNexis provides an option to search news by company ticker[129]. Using only news categorized by the ticker may exclude some relevant news items that are not recognized by the database's classifier. However this approach allows for significant reduction in the search results volume. In example, searching news for one week for Apple with the word '*apple'* in major world publications returns 1 171 news items that are mostly unrelated to the company. However, when using Apple's ticker, the news amount is reduced to 222 news items that are all related to the company. Hence, we choose to use this ticker information in our extraction process. We also restrict the sample by relevance score to exclude news that are not strongly linked to the company[130].

Second, when searching for a company, LexisNexis Academic provides the possibility to download the news in HTML format. The search is; however, limited to 500 news items per download. Due to this restriction, it is impossible to download all the news that we are

[129] SmartIndexing is a rule-based automated classification system that analyzes and tags online documents for relevant subjects, industries, companies, organizations, people and places. Researchers build searches using index terms to efficiently and accurately pinpoint results.

[130] The relevance score excludes news that refer to the company only few times. These weak stories may typically include a market, or an industry, overview where multiple companies are briefly mentioned. Such information may be relevant for the company; however, when reading the news item with a machine, a sentiment scoring system may not be able to pick the relevant message. Therefore, we exclude such news. LexisNexis categorizes the relevancy as follows: 90% to 99% is a major reference; 80% to 89% is a strong passing Reference; and 50% to 79% is a weak passing reference. We remove all messages under the threshold of 80%. By using relevance score, our filtering is similar to that of Tetlock et al., (2008) who examine each story's relevance to a company by checking whether the official and unofficial name of the company appear in the news item sufficiently many times.

interested in at once. Rather a search query has to be split into smaller pieces: each downloading less than 500 news items at a time. The LexisNexis service has only a web user interface: i.e., it is not possible to download news items without going through the website, manually navigating the web page. As downloading all of the >500 000 news would require more than thousand manual downloads, we develop a web scraper[131] that downloads the news from a list of tickers for a chosen time period. The scraper is also limited by the same 500 news item restriction as a human user. Therefore, we need to estimate a time-interval for each company that does not yield over 500 results for a query, so that the scraper is able to retrieve all relevant news, and is not missing out on news due to the maximum download restriction (e.g. for a search result yielding 600 results, we would miss the items 501-600 from our download). Therefore, we manually download a small number of news for each company to estimate a correct time-interval for each company that results in a suitable news item download amount. From there on, we let the scraper download the news for us. The details of how our scraper functions and navigates across the LexisNexis website can be found in Appendix F – Details on gathered media data, in sub-section: ‘*Details of the web scraper*’.

4.4.3 Data pre-processing

As an output of our scraper's downloads, we receive multiple firm-specific HTML-files, with 1 to 500 news items in each. Next, we need to convert the dataset into an easily usable format in a database; identifying each message, the corresponding date, company, publication and other similar metadata. The format in which LexisNexis provides the data does not follow standard formatting. Thus, sorting the messages appropriately requires an algorithm of its own. The technicalities of this processing can be found in Appendix F – Details on gathered media data, in section Details of processing LexisNexis data.

Next, we preprocess the text data. We sort the texts and remove irrelevant parts so that we will be able to easily analyze the texts data later. We perform the following actions on the data:

1) We remove all tables as they are typically financial figures, or other numerical or otherwise descriptive data that typically contain very little information on sentiment (e.g., Loughran and McDonald, 2011).
2) We check whether any sources provided by LexisNexis need to be excluded for our study. We manually go through the 25 news sources with the highest news volume to

[131] On the use of web scraper with news sentiment, see also Das,2010

see the kind of information provide. We would exclude e.g. machine generated news, and news that only present numerical data. However, such sources are not present in the 25 sources we check that collectively represent 56% of the news volume for the total of 482 sources.

3) In some cases, the message and other metadata (e.g. date, publication, etc.) cannot easily be distinguished from each other: often the data is highly amorphous in LexisNexis. In such cases, the news items in question are ignored from the dataset.
4) Similarly to Tetlock et al., (2008), we decide to exclude news where the length is below a specific threshold. Tetlock et al., exclude news with less than 50 words, while we choose to ignore all news with less than 100 words. We estimate that a news item with less than 100 words will have only few sentences, and insignificant value to the sentiment. Moreover, only one negative sentence could turn a short news item into drastically negative, and hence add considerable noise to our daily sentiment aggregation.
5) We also remove all media items one week post a firm's index inclusion, or listing, to avoid the well documented index inclusion phenomena (e.g., Shleifer, 1986; Tetlock et al., 2008; Loughran and McDonald, 2011). Following web research on listing dates, we exclude MasterCard first on 5 June 2006 (listed on 25 May), Philip Morris on 26 Mar 2008 (spin-off from Atria on 17 Mar), and Visa on 28 Mar 2008 (listed on 18 Mar).
6) We remove duplicate messages: messages that have the same date, ticker, publication, heading and length.
7) We manually scan the data to ensure that companies have a news volume approximately corresponding to their size, and that the news volume is approximately equally distributed across time.

Out of the downloaded 552 578 news items, we end up with a subset of 474 030 items after preprocessing. For summary on the screening process, see Table 4.

Table 4: Number of documents during preprocessing

	# messages	% original
Downloaded messages	552 578	100%
Metadata in difficult format	-14 501	97%
Message <100 words	-41 082	90%
Messages before listing	-1 931	90%
Duplicates	-21 034	86%
After preprocessing	474 030	86%

While our preprocessing may remove some relevant stories, we believe that the reduction in noise outweighs the loss in relevance considerably. Descriptive statistics of the final media dataset can be found in the tables and figures below.

At first, we review at the publications that our sample consists of (Table 5). We recognize that some of the top publications may be missing, such as the Wall Street Journal. This may be for example due to copyright issues, but we don't consider this a critical flaw[132]: the wide coverage of other newspapers should covey the same information content in most cases.

Table 5: News volume by publication

Publication type	# of news	% of total
Newspaper	282 493	60%
Newsletter	52 362	11%
Magazine	43 963	9%
Web Publication	28 048	6%
N/A	27 900	6%
Newswire	24 558	5%
Papers	3 118	1%
Transcript	2 178	0%
Journal	1 607	0%
Abstract	66	0%
Other	6 737	1%

Publication	# of news	% of total
Financial Times	39 616	8%
Daily Deal / The Deal	23 958	5%
The New York Times	18 905	4%
The Globe and Mail	15 958	3%
The International Herald Tribune	14 869	3%
National Posts Financial Post	13 816	3%
Biotech Business Week	12 943	3%
Drug Week	12 613	3%
The Washington Post	11 678	2%
TECHWEB	11 233	2%
Other (472 publications)	298 441	63%

Next, we review the sample for news volume by company (Table 6). We notice that the sample may deviate slightly from the long-term average: banks have received significant

[132] Moreover, Tetlock et al., (2008) find that Wall Street Journal articles do not have a significant impact on sentiment. Other studies have found similar results, as discussed in Section 2.3.

coverage during the financial crisis. On the other hand, banks also represent 10-20% of market capitalization[133] of S&P100 constituents, so one would expect financial institutions to have a relatively strong representation in terms of news volume.

Table 6: News volume by company

Company	# of news	% of total
Citi	33 444	7%
Goldman Sachs	32 870	7%
Google	30 602	6%
Ford	28 911	6%
Apple	26 581	6%
Boeing	20 513	4%
Microsoft	18 115	4%
J.P.Morgan	13 859	3%
Bank of America	12 781	3%
AT&T	9 908	2%
Other (90 companies)	246 446	52%

Industry	# of news	% of total
Banking	78 545	17%
Business Services	56 696	12%
Computers	44 766	9%
Trading	40 436	9%
Pharmaceutical Products	34 641	7%
Automobiles and Trucks	29 531	6%
Telecommunications	28 911	6%
Aircraft	28 055	6%
Petroleum and Gas	21 756	5%
Electronic Equipment	18 809	4%
Other (17 industries)	91 884	19%

Finally, we examine the timing of our news. We notice that the number of news decreases towards the end of our time period. This is likely due to the fact that news items are not updated or indexed immediately in the LexisNexis database, and thus some news or ticker metadata may not be saved to the database yet. We examine this trend for common patterns, but do not find systematic tendencies of missing data. [134]

[133] Fraction of market capitalization 17% on 1 Jan 2006, 12% on 31 Mar 2011.

[134] We study news volume per company but do not find systematic discrepancies between companies. Also, we test whether or not the data has a systematic change in terms of news publications by having a look at the number of different publications and the news volume of top publications over time. The number of daily publications remains relatively constant, though some individual publications have less news towards today's date. In conclusion, we do not find a systematic bias.

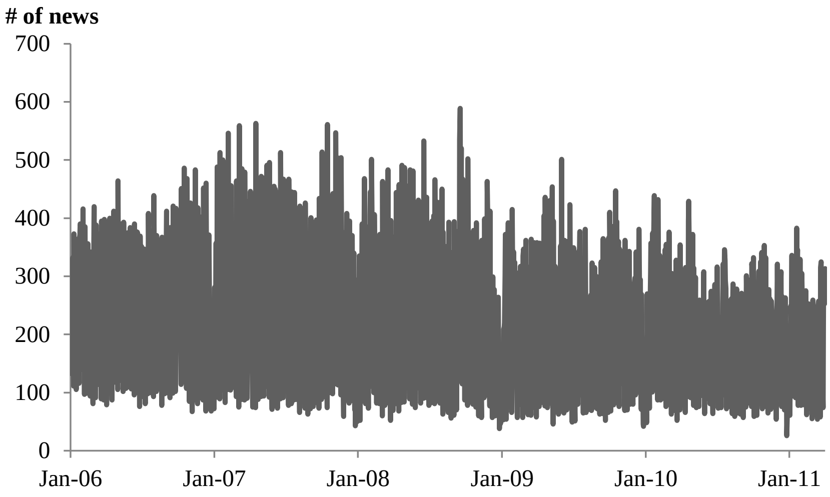


Figure 4: News volume by date

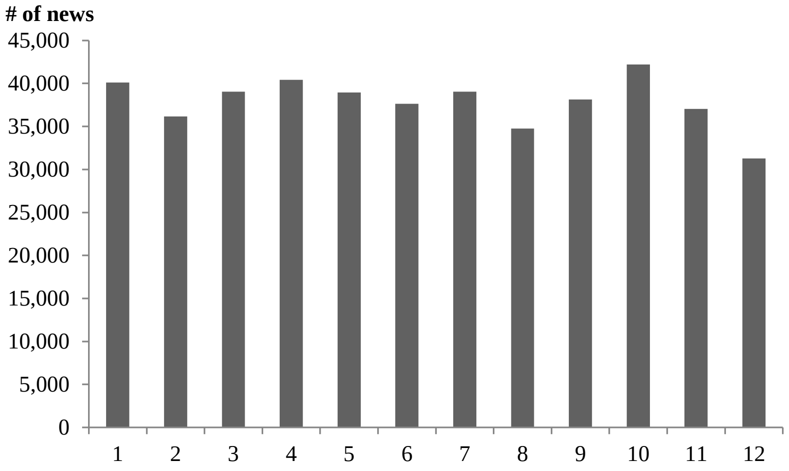


Figure 5: News volume by month[135]

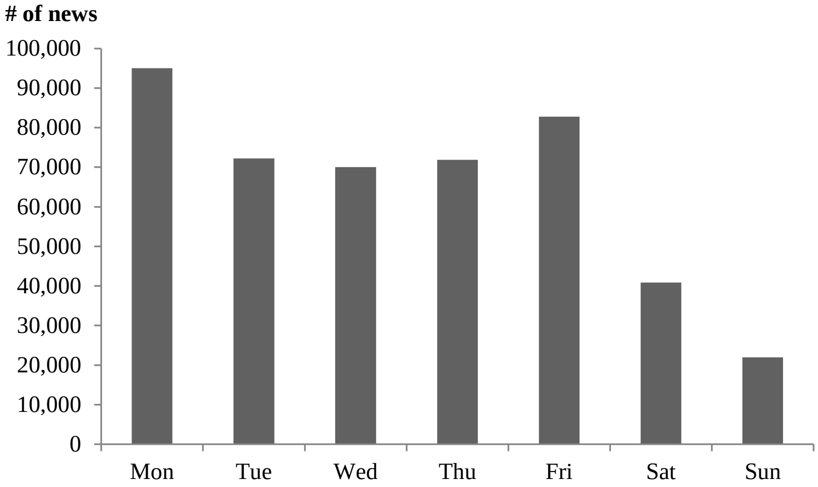


Figure 6: News volume by weekday

[135] News volume by month excludes 2011 data where we would have only Q1 available.

We conclude from the summary measures that our sample is sufficiently representative of common understanding on how news volume is distributed. With samples of financial and qualitative data, we move on to describe our research methodology.

5 METHODOLOGY

This section will discuss in detail the choices we have made when selecting methodology, and the potential drawbacks as well as the corresponding robust checks we have made to ensure that our choice of methodology is reliable.

First, we will describe in detail how we estimate investor sentiment. Second, we explain how we aggregate sentence and word scores into a firm specific daily sentiment score. Third, we introduce a number of alternative specifications for sentiment. Finally, we will move on to discuss how we study the relationship between our main independent variables and our dependent variables. The reader should refer to section 4.2 for a detailed discussion on the variables used in the study.

5.1 Estimating investor sentiment

Comprehending a text, understanding in what kind of a tone it is written (positive or a negative) is not an easy task. Different readers may interpret texts in different ways: this may depend on the context, the person's previous knowledge on the subject, and so on. In this section we describe the advanced methods for understanding the tone of a text. In our study, we calculate multiple scores for company sentiment, with the purpose that we can compare the effectiveness of different investor sentiment estimates as predictors of financial metrics.

First, we describe Quasi-compositional Sentiment Sequencing and Compression used by Moilanen et al., (2010) that has been shown to estimate sentence level sentiment more accurately than bag-of-words method. Similarly, Engelberg (2008) has used a methodology that analyses sentiment in sentence level in financial context called: typed dependency parsing.

Next, we describe our primary method for estimating investor sentiment in this study: the Linearized Phrase-Structure -model. Finally, we explain how we compile a set of new word-lists, and a training set, that are necessary for making the Linearized Phrase-Structure -model operational.

5.1.1 Quasi-compositional Sentiment Sequencing and Compression (MPQA)

Quasi-compositional sentiment sequencing[136] is a method proposed by Moilanen et al. (2010) for classifying sentences based on their sentiment polarity. The idea behind the method is that polarity in sentences typically follows similar patterns. For example, negative sentences have a certain pattern of negative and positive expressions in them. Once a reader has read a certain number of news, it is possible to use machine learning to find the patterns that make a reader categorize a news item as positive or negative. These patterns can then be used to categorize more similar news. This approach was used, among others, by Moilanen et al. (2010) to test classifying sentiment in sentence databases (e.g. with news headlines and financial text snippets).

For example, we can have a look at two following sentences (wordlist categories below).

The company	Is	doing	well.
(neutral)	(neutral)	(neutral)	(positive)

The company	Is	not	doing	well.
(neutral)	(neutral)	(reversal)	(neutral)	(positive)

A simple word count would notice in both sentences the word '*well*' and would likely annotate both sentences as positive. Quasi compositional sequencing, on the other hand, would notice the patterns 'neutral-*positive*' and 'neutral-*reversal-neutral-positive*', and hence, ignoring neutral expressions, could annotate the first sentence as positive and the second as negative.

5.1.2 Linearized Phrase-Structure -model

In this section, we present the Linearized Phrase-Structure (LPS) model for predicting semantic orientations of short economic texts. The Linearized Phrase-Structure -model is based on the Quasi-compositional Sentiment Sequencing method. However, we have improved and modified the method for financial context

[136] WE refer occasionally to Quasi-compositional Sentiment Sequencing with "MPQA". MPQA refers to the dictionary used by Moilanen et al. (2010)

In order to operate in financial and economic domains, a method should be able to recognize financial domain concepts and identify their semantic orientation based on sentence structure and domain knowledge. In addition to the domain specific aspects, a model should also have an ability to resolve conflicting cases where a sentence contains several semantic orientations, and be easy to retrain based on the feedback given by the users. To accommodate these requirements, the LPS-model is constructed in three stages: (i) identification of entities with semantic orientation; (ii) phrase-structure projection step; and (iii) multi-label classification step. In a nutshell, the model works as follows (for details and more formal definition of LPS, we refer the reader to Malo et al., 2013b).

Identification of entities with semantic orientation

Our model starts by identifying entities in a text stream. For example, this could mean that our algorithm detects the word “good” in a sentence with otherwise neutral words. To support recognizing which parts relate to each other in a sentence, we detect the phrase-structure[137] information to support us in understanding which parts in the sentence relate to each other. Once the phrase structure has been detected, we use entity recognizers to locate various lexicon entries (see section 5.1.3 for categories of lexicon entries) in the sentence. Finding the phrase structure and detecting the lexicon entries is illustrated as “Step 1” in Figure 8.

Once the initial set of entities has been recognized, heuristic rules are applied to merge neutral entities and to take into account the effects of polarity influencers (increase / decrease verbs) on the semantic orientations of other entities. The clear benefit of using such heuristics at this stage is that the number of entities is significantly reduced and the information value of the retained entities is higher. The following rules for entity-pruning are considered:

(i) Merge-neutrals-rule: If several neutral entities occur in a sequence, they can be combined into a single neutral entity which spans a large part of the given phrase. For example, when taken out of its context, the sentence “Net profit in the period in 2009 was EUR 29 million" can be considered to be a single neutral entity, since the default prior-polarity of "Net profit" is neutral and there are no polarity influencers in the sentence.

[137] Identifying phrase structures is a commonly used technique in natural language processing. Showing a sentence according to its phrase structures splits sentences according to parts such as noun phrases (NP), verb phrases (VP), prepositional phrases (PP), etc.

(ii) Polarity-influence-rule: When a phrase contains both a polarity influencer[138] and another entity whose polarity is modified by the influencer, we apply the influencer directly and retain only the main entity with modified polarity. For example, consider the two phrases in Figure 7, where in both cases we find a directionality which modifies the polarity of a financial concept. For sentence (a), we have "*EBIT*" as a financial entity whose polarity is modified by verb "*increase*" which leads to a positive overall orientation.

Instead of retaining both entities, we combine the entities by retaining EBIT and adjusting its prior-polarity from "*neutral*" to a modified polarity "*positive-up*" reflecting the fact that positiveness depends on the up-direction of events. The impacts of other influencers are accounted in similar manner. For instance, if a negator is attached to an entity with "*negative*" prior-polarity, we modify the polarity into "*negative-reverse*" to signal the fact that the entity has been merged with a polarity influencer.

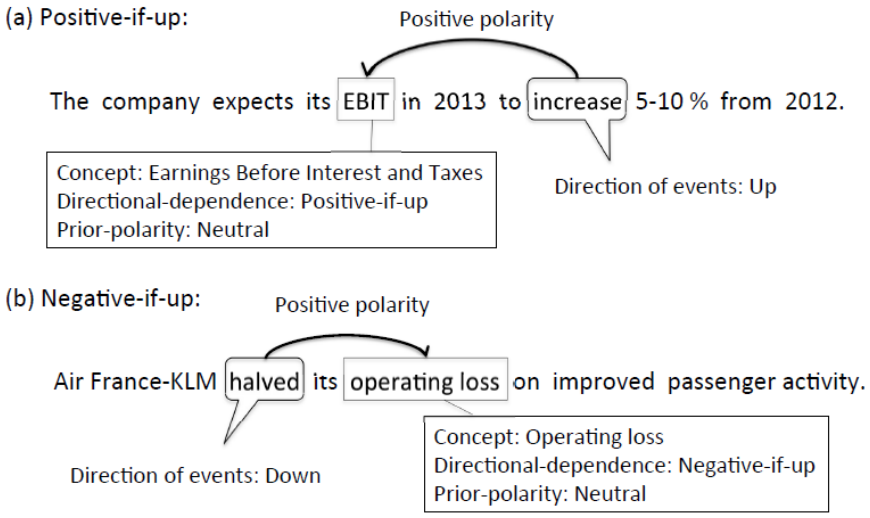

(a) Positive-if-up:
Positive polarity
The company expects its EBIT in 2013 to increase 5-10 % from 2012.
Concept: Earnings Before Interest and Taxes
Directional-dependence: Positive-if-up
Prior-polarity: Neutral
Direction of events: Up
(b) Negative-if-up:
Positive polarity
Air France-KLM halved its operating loss on improved passenger activity.
Direction of events: Down
Concept: Operating loss
Directional-dependence: Negative-if-up
Prior-polarity: Neutral

Figure 7: Finding polarities with Financial entities and Polarity influencers

To apply heuristics in a sensible manner, we use the detected phrase structure and only combine words within a restricted window around each entity. This process is illustrated as "Step 2" in Figure 8.

[138] Polarity influencers are described in the next section

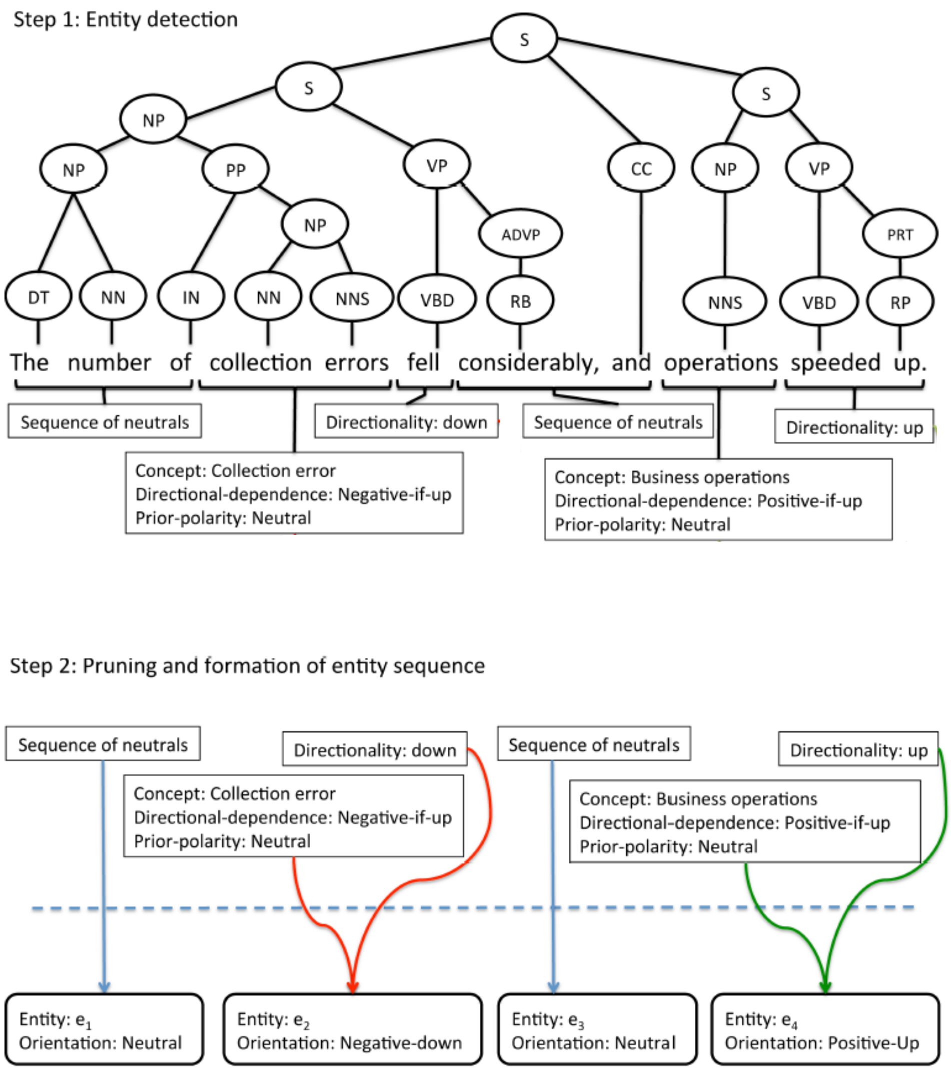


Figure 8: Identification of entities with semantic orientation with LPS

Our learning algorithm would be able to separately process sentences without these simplification (considering e.g. a sequence "*neutral*"-"*positive-if-up*"-"*positive*" and "*neutral*"-"*positive*" as different sequences). Thus, this merging may lead to a minor loss of accuracy[139]. However, this loss is likely significantly offset by the corresponding gain in computational speed and a smaller sample of training set required after merging.

[139] When applying the pruning rules, we do not fully hide the impact of polarity influencer. Instead of using the main semantic orientations "positive, neutral, negative", we distinguish the impact of pruning by using a "prior-polarity" information (e.g. "positive-up" instead of "positive"). This distinction is primarily motivated by added flexibility in the learning stage. Given the fact that heuristics are always bound to introduce some added uncertainty, we want to provide the model this way an opportunity to correct for mistakes and give the modified entities a differential treatment while making judgments on the overall polarity of the phrases.

Phrase-structure projection

In the second step we create a phrase-structure projection. The purpose of this step is to convert a sequence, e.g. "*positive-negative-neutral-positive*" into a formal presentation that can be used by our learning algorithm in the following Multi-label classification . Each projected sequence has an interpretation as a representative of an equivalence class of phrases with similar features, and the lengths vary depending on the complexity of the underlying phrase-structure.

To illustrate how this step works, assume that we had only three entity-types: positive, neutral, and negative. Then we can choose a coding where $\tilde{e}_+ = (1; 0; 0)$ indicates that entity is positive, $\tilde{e}_n = (0; 1; 0)$ indicates that entity is neutral, and finally $\tilde{e}_- = (0; 0; 1)$ implies that entity is negative. According to this system, if a phrase s has an entity-sequence with categories positive-positive-neutral-positive, i.e. $\tilde{e}_+ \, \tilde{e}_+ \, \tilde{e}_n \, \tilde{e}_+$, we can write

$$\text{Sequence} = (1; 0; 0; 1; 0; 0; 0; 1; 0; 1; 0; 0; \ldots)$$

as a bit-sequence representation of the given equivalence class, where components beyond 12th are all zeros. A presentation of this kind can then be presented to a linear multi-label classifier, which learns to associate the sequences with corresponding semantic orientations indicated by the annotators.

As the phrase-structure projection step is in essence a technical step to convert data into a form that can be used in the following classification step, we have left further details out of this paper. For more formal definitions of this step, we refer the reader to Malo et al., 2013b.

Multi-label classification

The final step, multi-label classification, aims to classify the sentences into different polarity classes based on the entities recognized. For this, we require a learning mechanism that is able to form rules from a sample of annotated sentences. This learning mechanism should (i) be able to handle large-dimensional feature spaces in an effective manner; and (ii) be able to perform multi-label classification. After a few preliminary experiments, we decided to choose the a multi-class SVM[140] approach with "*one-against-one*" strategy, which has shown good performance in comparison to alternatives based on "*all-together*" methods or "*one-against-all*" and DAGSVM strategies. The choice is also well supported by the study of Hsu and Lin

[140] For use of SVMs in previous literature, refer to section 2.3.1

(2002), who evaluated a number of alternative multi-class SVM models with different estimation strategies in the light of large-scale problems.

5.1.3 Lexicon entries for Linearized Phrase-Structure -model

When detecting sentiment in text, not all words are equal. For example, adjectives often are in a more significant role than prepositions. Looking only at a certain number of words is not always enough to detect the correct sentiment, but limiting our analysis to certain words can significantly reduce the requirements for detecting sentiment, as analysis of the meaning of the sentence on multiple levels is not be required. The purpose of a lexicon is to define the categories of words that we consider when analyzing text. Our lexicon has multiple categories, and any word in a certain category is considered with equivalent weight. Our lexicon consist of categories for words that are Positive, Negative, Negation words, Financial entities that turn into positive or negative if combined with certain words, and words that can impact the polarity of financial entities.

Using the more general MPQA lexicon by Wiebe et al.(2005) and the financial polarity-lexicon by Loughran and McDonald (2011) as a starting point, we propose the following modifications that infuse further domain-specific knowledge into the sentiment models: (1) addition of domain-specific concepts which can influence the overall semantic orientation of a sentence; (2) addition of verbs and expressions which help to detect the direction of events (e.g. whether the profit is expected to increase or decrease); (3) addition of information on how the polarity of different concepts depends on the expected direction of events (e.g. result is positive when it is expected to increase, but neutral or negative when declining). The used wordlists are described below.

Positive and negative words

Most typically sentiment has been detected from lists of positives, negatives and negation words. To create such word lists, we build on top of the lexicon derived from Multi-perspective Question Answering (MPQA) corpus of opinion annotations; Wilson (2008). The lexicon consists of only single-word clues, and each entry of the lexicon is equipped with information on degree of polarity (positive, negative, neutral), subjectivity, the word’s lemma form and the default part-of-speech the word has.

By utilizing these general entries as a seed for our lexicon, we obtain a good coverage of the most commonly encountered subjective expressions. However, as mentioned earlier, the general polarity lexicons are not directly applicable to the financial domain, since many commonly used expressions may take a different meaning in the financial and economic context. To accommodate the domain specific requirements, we have augmented the MPQA-based dictionary with the finance-specific lists compiled by Loughran and McDonald (2011). When overlaps were encountered while merging the lists, the financial domain sentiment was preferred over general prior-polarities specified by MPQA.

Financial entities

On top of the traditional dictionary categories, we add words that, combined with a verb indicating movement up or down, will have an impact on sentiment. For example, '*sales grew*' has a positive sentiment, while the word '*sales*' by itself has no sentiment. To establish such a word list in the financial context, we download the financial online dictionary Investopedia.[141] Through Investopedia dictionary, we find a list of 16,178 different financial concepts. From the list, we remove all words that have only one or two characters. To zero in on the most important terms, we take a random sample of 100,000 news articles from our sample, and count the occurrences of all financial terms in the news. We find occurrences of 4,389 of the different dictionary terms in the news sample, and order the terms according to frequency. We go manually through all the terms that have more than 200 occurrences in our sample, which includes 684 terms. Next, we remove words that are not only financial concepts but also common English words such as: "SPAN" abbreviation for '*standardized portfolio analysis of risk'* in the financial context. After that, we start removing words that have no meaning, or the meaning is unclear to the sentiment of a company: e.g. '*shares*', '*EUR*', '*plc.*', etc. While the words that have over 200 occurrences in the sample represent only a fraction of all the words (~16%), their volume represents approximately 94% of all financial terms in the text. Thus looking through the remaining 84% of the terms would only increase the recognized terms by 6%, and thus we ignore the 84% of terms from our list. A list of the selected words can be found in Appendix H – Financial entities.

Out of the sample of 684 terms that we manually went through, 51 terms have in our view a very clear effect on sentiment, and 177 have an effect on sentiment in most cases when combined with a verb representing movement up or down. As most of the time the aforesaid terms will yield correct interpretation of true meaning, we merge the terms in order to create two word lists:

[141] While Investopedia may not be the most prestigious dictionary, we select it due to ease of access: a free dictionary in electronic format, and a very wide coverage of concepts

- ‘*Positive-if-up*’: e.g. “EBIT”
- ‘*Negative-if-up’*: e.g. “taxes”

Polarity influencers

In addition to words that have a polarity directly attached to them, there are a number of other factors which can influence the overall semantic orientation of a phrase. This broad class of operators, which can modify the contextual semantic orientation, is generally referred to as polarity influencers. The most commonly encountered polarity influencers tend to fall into one of the following categories:

- Negation words. After going through movement linking words and corresponding movement verbs, we download words that can reverse the meaning of a sentimental word: for example: ‘*It is going well.*’ vis-à-vis ‘*It is* <u>not</u> *going well.*’ To do this, we download the General Inquirer category: negate. Similarly to the verbs before, we remove words with different financial meaning. Removed words from negate-list include: ‘*account*’, ‘*uncertain*’, and ‘*uncertainty*’, ‘*unemployment*’ and ‘*vice*’.
- Boosters and diminishers. Another important class of polarity influencers consists of words which can intensify or reduce the degree of positiveness or negativeness of an expression; e.g. “paying off the national debt will be ‘*extremely painful'*" or “*little threat*".
- Modal operators. A modal operator is a verb which modifies another verb and describes the “*mode of operating*" by setting up a context of possibility or necessity; e.g. *“we have to revise the policy*" or “*we can revise the policy*".

Furthermore, to accompany the financial entities -word lists, we must identify words that describe movement up or down. Directionalities, described below, are an additional category of polarity influencers that we have used.

Directionalities

Directionalities refer to words that describe movement, e.g. “*increase*”. Harvard’s psychological dictionaries have been commonly used in sentence tagging, and we download Harvard IV word lists from the General Inquirer. We create two wordlists from the following General Inquirer categories:

- ‘*Increase*’ - Harvard dictionary categories: increase and rise
- ‘*Decrease’*- Harvard dictionary categories: decrease and fall

Additionally, we go through the categories to see if they include words that would have a different meaning in a financial context. We find, and remove, the following words that tend to have a different meaning in a financial context: '*inflation*', '*people*': removed from increase-list, '*discount*', '*recession*': removed from decrease-list.

Finally, we combine our word lists mentioned above to end up with a final set of 11 word lists. As Linearized Phrase-Structure -model identifies lemmas for words, we included each word in the list only in their lemma form. The lists are described in Table 7: Wordlists for Linearized Phrase-Structure -model.

Table 7: Wordlists for Linearized Phrase-Structure -model

Wordlist	Number of words in list
Positive	264
Negative	1202
Negate	202
Increase	117
Decrease	111
Positive if up	121
Negative if up	56
Boosters	85
Diminishers	92
Modal Words – Strong	12
Modal Words – Weak	15

5.1.4 Training set for Linearized Phrase-Structure -model

The goal of a financial sentiment training set is to train an algorithm to take as input all available media items and to create an estimate of investor sentiment score that is equivalent to a sentiment score given by an analyst after reading the same data. In fact, if the algorithm is sufficiently developed, it could yield even better results than the analysis of one analyst. Indeed, it should, in theory, be able to match the aggregate sentiment score of several qualified analysts with access to the same data.

In order to correctly identify polarity on the sentence-level, a computer would need to know both the sentiment of words (word lists), and the sentiment of sentences: how do polarized words interact when combined on the most elemental levels. While it is clear that a sentence with 3 positive words and 1 neutral word is likely a positive sentence, it is much harder to identify what the sentiment is in a sentence with 2 positive and 2 negative words. Our training set aims to create a sample of different patterns and their outcomes that can be used for classifier training.

Training set sample

As discussed by Loughran and McDonald (2011), it is well understood that the vocabulary and expressions used to describe economic events and company related news are not identical across media. To build and evaluate models which are dedicated for capturing semantic orientations in economic texts, it is important that the training material provides a good coverage of the commonly used domain-specific expressions.

Until now, very limited efforts have been taken to build corpora which cover economic or financial domains, and to the best of our knowledge, none of the existing datasets provides phrase-level annotations for news documents. Furthermore, many of the data-banks mentioned in the literature are known to be reserved for proprietary use only; e.g. O'Hare et al. (2009). Therefore, to alleviate the data gap, we will now briefly outline our financial news-phrase dataset, which can be used as a gold standard for evaluating the performance of sentiment models dedicated for economic texts.

The sample for the training classifier is picked from full sample of articles that we have downloaded. Out of these articles, we select a random subset of 10,000 articles, with weighted probability of including each sentence into the sample so that we even the distribution of:

- Small and large companies
- Companies in different industries
- News sources

As described earlier on preprocessing of our media sample, we remove also in this case items such as html-code and tables from the selected articles. For the training sample of 10,000 articles, we look for words in our polarity lexicon in each sentence, and select only sentences that include words with polarity (see e.g. Maks and Vossen (2010)). This reduces our sample to 53,400 sentences: each having at least one recognized word from our word list, or a combination of recognized words.

Ideally a sentiment-recognizing algorithm should be able to take any recognized word combination and assign a sentiment score for it. However, the number of different combinations that we find in our sample is 13,184, and we would need a large sample of annotated sentences for each combination to reach a good confidence level for the algorithm. Furthermore, most of the combinations appear only once or twice in our sample of 10,000 articles, and would add negligible value for the algorithm. Thus, we choose to rank the different combinations according to frequency distribution so that we can focus on the most salient patterns. For example, the most common patterns cover 9% of all of our sentences.

After ranking, we start choosing samples of approximately 30 sentences from each pattern combination, starting from the most frequent combinations. We iterate our sample during the annotation process: depending on the results, we may increase the frequency of sentences in a category, or stay with 30 sentences. For example, if a vast majority of the sentences in the same category receives the same annotation - for instance, 29/30 sentences are positive for a given sequence - we conclude that the category is correctly processed. On the other hand, if there is large deviations in the classified sentiment for a certain combination: e.g., 50-50 split etc., we increase the number of unique sentences annotated for the category at hand to ensure we get a better understanding of how the sequence should be annotated.

Annotation labels

The most evident polarity to detect is positive or negative. For example, Maks & Vossen (2010) use tags ‘*pos*’ and ‘*neg*’ to denote this. Other commonly used metrics include:

- Subjectivity vs. objectivity: objective truth vis-à-vis an opinion of the writer (e.g. Maks & Vossen, 2010)
- Whose opinion is it in the sentence: authors or someone else’s (e.g. Maks & Vossen, 2010)
- Relevance: is a sentence relevant for the topic, or not.[142]

Differences in sentiment arise depending on the interpreter’s background, purpose and own views. In common language: ‘*Company made a loss of 10 EURm*’, would be negative. A financial analyst on the other hand, would need more information: What did the company make last year? What did the market expect? Was the loss just a reflection of a ramp-up cost of a long-term investment?

[142] For example, Hsueh et al. (2009) start the annotation process by tagging sentences that are irrelevant to the political candidate that they are studying

The question we wish to answer in financial context is whether or not a sentence is positive, neutral, or negative. To illustrate:

(1) The headquarters of this company looked beautiful in my opinion.

(2) The company doubled its profits this year.

To make sure we will tag sentences like (2) as positive, and preferably ignore sentences like (1)[143], we instruct our annotator to categorize sentences based on the impact on company's share price. We also provided a set of example sentences that we annotate ourselves for reference to the annotator. After the aforementioned, we ask her to use a 9-step scale of labels to annotate where she expects the share price would move following the publication of the news item. For a description of our instructions and the used labels, see Appendix G – Annotation instructions.

The above demonstrates the two goals we have: first, we aim to define if a sentence is positive or negative for the company - from an investor point of view; second, we wish to take into account that there may be difference in the level of polarity of a sentence. Similarly to Wiebe et al. (2005), we follow three principles with the guidelines we have established for our annotation process:

- ❖ There are no fixed rules about how particular words should be annotated. The instructions describe the annotations of specific examples, but do not state that specific words should always be annotated a certain way.
- ❖ Sentences should be interpreted with respect to the context. The annotators should not take sentences out of context, start speculating on their prior knowledge on the company, or think what the sentences could mean, but rather should judge them as they can be interpreted in isolation.
- ❖ The annotators should be as consistent as they can be with respect to their own annotations, and the sample annotations given to them for training.

We do not annotate for the following: the opinion holder, or relevance of the sentence, as we assume that the aforementioned will play a smaller role in the sentiment. Furthermore, we assume that most sentences should be relevant for the company for two reasons: first, they are

[143] We recognize that sentence (1) may also have some value, and it may also be good to annotate this as positive. For more on this, refer to our training set B at the end of this section.

parts of articles that we classified previously as relevant to the company, second: they are mainly texts from financial press that should be concise and up to the point.[144]

Annotators and their training

In order to tag the sentences easily, we create an excel template where an annotator can easily select a number between 1 to 9 in order to select a tag with up or down arrow keys. This approach is chosen both for convenience, and to make annotating as fast as possible.

Our annotator is hired through an online service Elance.com where it is possible to hire freelance low-cost workers. Contrary to some other annotation studies, the approach was chosen over other alternatives such as crowd sourcing services (e.g. Hsueh et al., 2009 on using Mechanical Turk). We wanted to select an annotator with appropriate background and to be familiar with her to ensure that the quality of work would be consistent. Furthermore, we wanted to have direct lines of communication with the annotator as this can be beneficial when discussing complex cases. Compared to many crowdsourcing tasks, the nature of our task is different: requiring special expertise as people with no financial background may have misconceptions such as: *'lay-offs are always bad news for a company*', *'making (any kinds of) profits is always positive'*, etc.

We screened the service for low-cost workers who have a business background and strong English skills. Next, we asked a few promising candidates to annotate a small piece of sentences, and compared their work to annotations that we had done. The annotator applicant who was able to give sufficient quality (>85% correct answers for a set of 75 sentences) was chosen for the task. Our annotator is from the Philippines, had studied business administration, and had 5 years of experience in rewriting financial texts as a senior editor for different newspapers.

We start by annotating a small number of sentences ourselves in order to give the annotator a small set of readily annotated sentences that she could use as a benchmark for future annotations. Also, the annotator received the annotation instructions (Appendix G – Annotation instructions) explaining how the sentences should be tagged.

In addition to the initial instructions, the annotator asked us several questions in uncertain cases to ensure that she was tagging sentences in the expected way. Some necessary clarifications included:

[144] For the biases that result from this simplification, please refer to results in section 6.1.2.

- When does a stock go up/down a lot vs. a little: the difference between categories 7 and 9. While some linguistic expressions give clear indication of this, numerical expressions are more difficult. We arbitrarily defined that changes in key metrics (e.g. sales, profit/loss etc.) are small if <5%, and large if >10%[145]
- Situations where there are two contradicting statements in the same sentence, for example: *'Operating profit decreased while net turnover increased.'*. We agreed that these kinds of sentences would require a very high level of judgment to make correct conclusions, thus we instructed the annotator to annotate the sentences as '*Either way'*

To make sure that all of the sentences were tagged with good quality, we sent the sentences in batches to the annotator - 75 sentences; followed by: 1,000, 2,500 and finally 1,500. After each larger batch was completed, we would manually annotate a randomly selected 10% of the tagged sentences, and compare this with the annotation results. We used the previously said method to verify that the annotation quality stayed above 85%[146], we sent the annotator our own views of the control annotations with explanations, and the next batch. Thus, out of the set of ~5,000 sentences, we had checked ~500, and could verify that the quality had consistently stayed above 85% compared to our own annotations.

Furthermore, most of the errors in tagging were only 1 step away from the correct tag (see Table 8); for instance, positive had been tagged as neutral or vice versa, or negative as neutral or vice versa. More severe cases, such as tagging positive sentences as negative, were extremely rare.

Interannotator agreement

After the annotation process, we compared the annotations that had been annotated by both the annotator and in our control annotation. In Table 8 we show the annotation results, and similarly to other authors (e.g. Somasundaran et al., 2006), we calculate Cohen's (1960) Kappa measure[147].

[145] Later on in our study we merge all positive categories into one, and all negative categories into one. Thus, this has no impact ultimately on our results.

[146] When comparing positive, negative and neutral findings – not the degree of positivity as this would be often a more subjective view

[147] Cohen's Kappa measures how much annotators agree compared to how much they would agree by chance. This is a measure that gives us a value of $\kappa = 1$ if annotators were in complete agreement, and $\kappa = 0$ if there is no agreement other than chance.

Table 8: Interannotator agreement (Training set A)

		Full-time annotator								
		+ + +	+ +	+	n1	n2	n3	-	- -	- - -
Control Annotation	+ + +	26	-	-	-	-	-	-	-	-
	+ +	**2**	63	-	-	-	**2**	-	-	**1**
	+	-	-	4	-	-	-	-	-	-
	n1	-	**16**	-	115	-	**2**	-	**1**	-
	n2	**1**	-	-	-	144	-	-	-	**1**
	n3	-	-	-	-	-	-	-	-	-
	-	-	-	-	1	-	-	4	-	**1**
	- -	**1**	-	-	-	-	-	-	9	**1**
	- - -	-	-	-	-	-	-	-	-	15

+++ = Up a lot, --- Down a lot, n = neutral

κ=0.90

In total, we reach an annotator agreement of 93% with 380 sentences. In particular when simplifying the categories to negative, positive and neutral, the interannotator agreement further improves to 94% (see Table 9).

Table 9: Interannotator agreement with simplified categories (Training set A)

The interannotator agreement from above is presented below, after simplifying the number of categories to three

		Full-time annotator		
		Positive	Neutral	Negative
Control Annotation	Positive	95 *23 %*	**2** *0 %*	**1** *0 %*
	Neutral	**17** *4 %*	261 *64 %*	**2** *0 %*
	Negative	**1** *0 %*	**1** *0 %*	30 *7 %*

κ=0.88

Out of all interannotator disagreements (in Table 8), 16/30 relate to situations where the full-time annotator considers a company reporting profits, sales etc. as a positive event (e.g. "*In 2005 the bank posted a net profit of 8.2 EURm .*"). As the company may post a profit below expectations, these sentences should rather be annotated as neutral in our view. We instruct the annotator during the annotation process on this, and thus avoid the same disagreement towards the rest of the annotation process. Apart from aforementioned, there appears to be no systematic error that the main annotator is making. Some disagreement cases are sentences where the sentences are complex and a reader needs to read the sentence multiple times to ensure understanding. We also identify sentences that present a view from two angles as one source of interannotator disagreement. Additionally, some financial expressions may also be difficult to interpret, and may depend on the context whether they are positive or negative. For example '*dividend cuts*' may be seen as a negative signal, unless accompanied with a good reasoning (see also Mitra and Mitra, 2010). Finally writing styles such as cynicism, (Hsueh et al, 2009), and inherent variability in a word's meaning (Maks and Vossen, 2010), may be sources of interannotator disagreement.

Maks and Vossen (2010) use a third annotator to get to a '*gold standard,*' and thus we also consider the possibility of using further annotators. Hsueh et al. (2009), on the other hand, conclude that it is possible to use only one expert annotator, finding that this gives 97.4% correlation to the gold standard in their case. When comparing to similar annotation studies, our Kappas compare relatively well (cf. e.g. Maks and Vossen, 2010, κ=0.80). As our algorithm will make its sentiment estimates based on probabilities, and use at least 30 labeled sentences for each pattern, a 100% agreement will not be necessary for the algorithm to work correctly. Consequently, we do not see it necessary to add more annotators[148]. However, we notice after further analysis (see section 6.1.2) that a somewhat different training set may be more ideal for the purpose of training our algorithm.

Training set B

After our original training set, we have an excellent set of tagged sentences that represent sentiment in different sentences. This training set is in our view a solid benchmark to see whether an algorithm is classifying sentences correctly. However, our algorithm can only recognize the sentiment, but is unable to recognize the credibility of the author (Appendix J - Error descriptions for LPS: Company talking in advertising like -tone about its' own operations) or the relevance of a sentence for a company's success (Appendix J - Error descriptions for LPS: Inability to recognize significance of events, Positive convention of talking about something, Inability to understand the magnitude or value of items). A training set where the '*correct answer'* includes also deductions based on this information has more correct annotations, but they also include more noise from the algorithm: our approach does not include a model for assessing relevance or credibility, and therefore sentences where an annotator considers this information effectively increase noise. To adjust our training set for

[148] The same set of sentences have been further annotated by more annotators in our parallel study. For details, see Malo et al. (2013b). However, even after further annotations, the training set remains in line with our original annotator's categorizations.

better results, we create a second training set (Training set B[149]) that is annotated by a researcher with a business background from Aalto School of Business. Thus, we have two different versions of the same training set:

A. Training set with credibility and relevance assessment: a person with a financial background reads the sentences and uses all their knowledge, except for company-specific knowledge, to annotate the sentences
B. Training set without credibility and relevance assessment: as above, but the person does not assess the credibility or relevance of the sentence.

The difference between the two training sets can be characterized with the three example sentences below in Table 13:

Table 10: Differences between training set A and B

Sentence	Training set A	Training set B	Difference
"I think my company will beat its competitors" -CEO.	Slightly positive / neutral	Very positive	Credibility of author
The company's 100th year celebration party was a great success.	Slightly positive / neutral	Very positive	Relevance of adjustment
"The company will likely beat its competitors" - Financial Times.	Very positive	Very positive	N/A

It is clear that Training set A is closer to a '*true*' sentiment estimate, and that the annotations are superior to Training set B in this sense. Credibility and relevance are; however, usually not assessed from the polarized sentences. Rather, they require knowledge on the credibility of different sources, and on the relevance of different events to companies, etc. As this would require us to identify a much wider variety of objects in the sentences, and represent a number of studies of their own, we create Training set B so that we can directly relate polarized words to sentiment.

We create our training set B using the set A as a basis. However, we especially go through the neutral sentences in order to detect cases where a sentence has been classified as neutral due to lack of credibility (e.g. a biased source, such as CEO explaining how good his own products are) or lack of relevance (e.g. an event that is considered unimportant). For such cases, we annotate the sentence with the polarity even if we know that this is likely not

[149] Training set B corresponds to Dataset I by Malo et al. (2013b), Training set A to Dataset II.

relevant for the stock price. To further simplify the exercise, we only tag sentences on a 3-step scale case in Training set B. The interannotator agreement between Training set A and Training set B is summarized in Table 11.

Table 11: Interannotator agreement between Training set A and Training set B

		Training set A		
		Positive	**Neutral**	**Negative**
Training set B	**Positive**	23 %	**15 %**	**0 %**
	Neutral	**2 %**	46 %	**0 %**
	Negative	**0 %**	**5 %**	9 %

κ=0.61

As expected, interannotator agreement is lower in this case compared to the control annotation, as the used instructions have been different. In particular, many sentences that the first, “stricter”, annotator characterizes as neutral, are categorized often with a sentiment by the second annotator. For example a sentence that is written in a positive tone but is not relevant for the company should become annotated as neutral in Training set A, but positive in Training set B.

5.2 Sentiment aggregation

In this sub-section, we will deal with several considerations that arise when calculating sentiment scores for full articles (as opposed to individual sentences), as well as when aggregating several sentiment scores from multiple articles on a given day into one sentiment score. The sub-section will proceed by first discussing the aggregation technique we have employed, and then move on to discuss additional considerations we have not yet explored when discussing our sentiment estimation methodology.

5.2.1 Daily sentiment aggregation

Once we have applied our different methods of investor sentiment estimation, we need to aggregate the polarized results for a document. In the case that there are several articles for a given company, we wish to further aggregate the sentiment scores of these articles within a day into an aggregate sentiment score for that day for a given firm.

To ensure that our sample articles are relevant, we start out by filtering news based on their characteristics. As explained in Section 4.4.3, we have already removed news with less than 100 words. In addition, when doing a word count, we require each news item to have a minimum of three negative words with two of them being unique (e.g., Tetlock et al., 2008). Similarly, we also require documents to have at least two sentences with negative polarity (our word count may still use these news items). The exclusion is done in order to eliminate stories that contain only tables, or lists, with mostly quantitative information. For example, a table might contain an individual word multiple times in the header of the table, and thus could add considerable noise to the sentiment if it were included in the sentiment score. The articles that meet the aforementioned criteria are included in our news sample and hence in our sentiment score.

To consolidate polarized elements, Das (2010) suggests calculating '*sents*', where a positive ('*BUY*') signal is calculated as +1, negative ('*SELL*') signal as -1, and a neutral ('*HOLD*') signal as 0. From the news items that are left for aggregation, we aggregate the sentiment score *Neg*[150] for document d as

$$Neg_d = \text{Number of negative words (sentences) / Total words (sentences)}$$

As can be seen, we use the same method of aggregation for word count and the Linearized Phrase-Structure -model. This is done to keep the consolidated sentiments consistent, which will allow us to better compare the methods.

Aggregation of the daily sentiment based on multiple articles with different scores can be done either by (a) using averages of the sentiment per article, or by (b) combining all articles within a day into a composite article. The chosen method can significantly impact the weight that each source gets in the sentiment score. In option a) each document can be set to have a weight (equal weight, or some other weight), while this is not possible if we aggregate all word and sentences directly into a composite article (option b). Previous studies have counted the aggregate daily media sentiment for a company by pooling all news items together in order to create a composite article. Then, negative words in all articles during respective day / total words in all articles during respective day (e.g., Tetlock et al., 2008; Engelberg, 2008) would reflect the sentiment of the day. However, we choose to differ from this approach. Consider a day when six articles are published: an article with 1,000 words, 100 of which are

[150] As negative news has been shown by previous literature to be most influential, we use in consolidated sentiment scores the negative sentiment.

negative; and five articles with 200 words, 0 of which are negative. With prior literature's method, the aforementioned example would yield a sentiment score of:

$$Neg = \frac{(100+0+0+0+0+0)}{(1000+200+200+200+200+200)} = 5\%$$

The weight of an article for the daily sentiment score would thus be directly proportional to its length which can simply be a function of writing style. While writing in a certain style may impact people's perception, we believe that multiple sources weigh more in the formation of an aggregate sentiment than the length, and negativity conveyed possibly by a single source[151]. Therefore, we aggregate daily sentiments based on equal weights between sources; in other words, using an average of the articles' sentiment scores within a day:

$$Neg_t = \sum_{d=1}^{n} \frac{Number of negative\ words\ (sentences)}{Number of all\ words\ (sentences)} \times \frac{1}{N}$$

N= number of articles during day t

The roots of our approach can be traced back to behavioral finance theory[152]. According to mental accounting, people do not aggregate related information rationally but consider it in insulation. As discussed, aggregating using composite articles overweighs lengthy articles vis-à-vis the different number of articles. We hypothesize the following: agents do not aggregate different news during a day but use a 1/N style heuristic rule in forming their sentiment estimate for the day, leading to equal weighting of news: averaging.

5.2.2 Considerations on daily sentiment aggregation

Prior literature studies have estimated daily sentiment scores using different methods besides the simple fraction of negative words to total words. For instance, Tetlock (2007) uses standardization of negative fraction as follows:

$$Neg = \frac{\mathrm{Neg} - \mu\mathrm{neg}}{\sigma\mathrm{neg}}$$

where μ_{neg} is the mean of *Neg* over the previous 365 days and σ_{neg} is the standard deviation over the same period. Standardization might be needed, for instance, in the case that different

[151] In fact, proxies of impact should be the prestige and number of readers of a source in a more sophisticated sentiment algorithm.

[152] See Section 2.2. for more discussion on behavioral finance, and the related sources.

publications change their coverage style, or that some new publications have been added to a news database during the sample period. However we do not follow this approach as previous studies have not found a significant difference between standardization and the simple fraction (Tetlock et al, 2008; Engelberg, 2008). Also, it is possible that there is a justified shift in negativity during a sample period. For example, our sample reaches over the financial crisis. Therefore, it could be justified that the sentiment would change over time, and smoothing the sentiment with standardization would distort the correct sentiment.

Besides standardization, term-weighting has been used when estimating daily investor sentiment (Loughran and McDonald, 2011). The method takes into account the length of a document, the frequency of terms, and commonality of terms within the entire corpus. According to Loughran and McDonald, term weighting can be especially beneficial when using a dictionary that is not tailored for the context it is being used in: for example using the Harvard psychology dictionary in financial context. However, as we are using a context-specific dictionary, and wish to stay consistent with other studies: only Loughran and McDonald have used term-weighting, we refrain from using term-weighting.

As we are using closing prices, we need to take into account the sentiment changes occurring during weekends to have an accurate reflection of the relationship between financial metrics and sentiment. Therefore, we calculate the sentiment scores for Monday’s by adding the sentiment of the weekend to the sentiment of Monday. By doing so, we take into account that the change from Friday’s closing price to Monday’s closing price includes news from Saturday, Sunday and Monday. We use the aforementioned method also for other days when the stock market has been closed.

Once we have calculated these sentiment scores, we further check the scores for seasonality and industry trends to make sure that there is no systematic bias impacting the sentiment. We test whether there are time periods when negative news are more common, and if negative news are constantly reported more in a certain industry. However, as we previously noted when discussing the standardization possibility, there may be a good explanation why these trends are occurring: i.e. a certain industry could be constantly declining in value and therefore warrant a constant increase in negative sentiment towards it. However, we wish to make sure that there is a logical explanation behind such trends, and that it is not simply a matter of what news are being included in our sample.

5.2.3 Robustness checks for LPS-sentiment

We calculate the LPS sentiment for each company. To ensure this sentiment correctly represents the market sentiment, we manually go through the resulting sentiment for the whole market: measured by S&P100, and for a few selected companies and the banking industry. For a summary of these sentiments and their comparison with the corresponding stock price, or index, see the figures below[153].

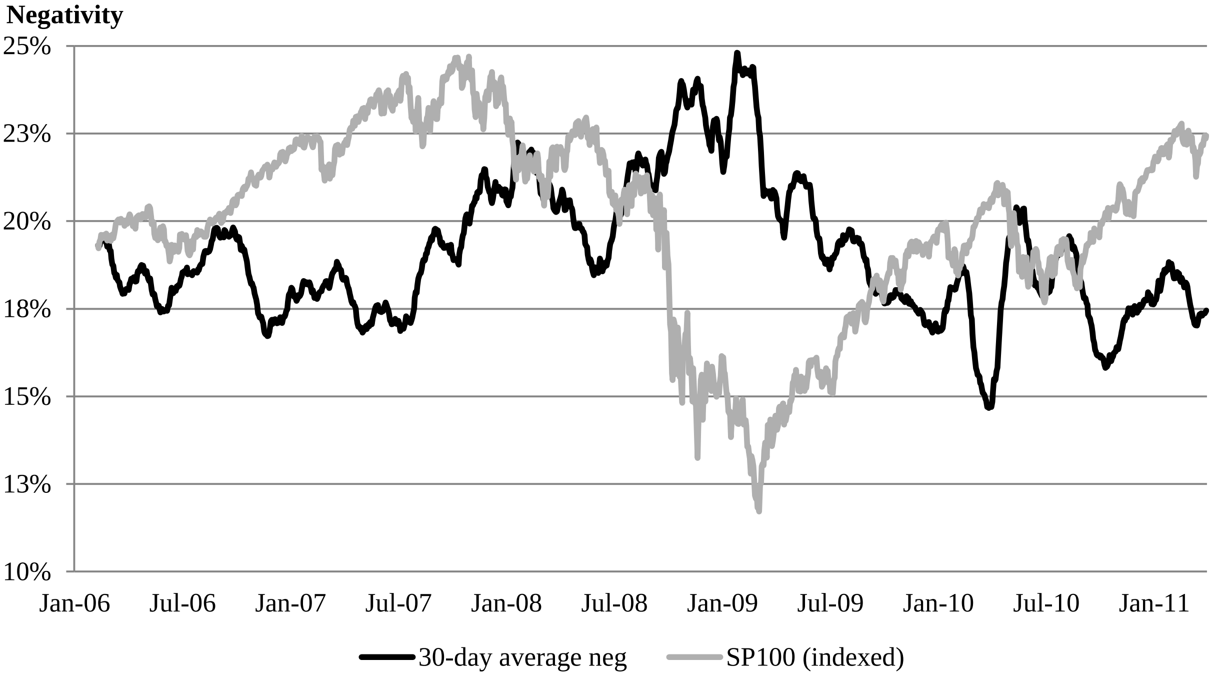


Figure 9: SP100 sentiment and stock index

[153] We show 30-day average sentiment, as a shorter term sentiment would be difficult to display due to significant volatility in the daily sentiment metric.

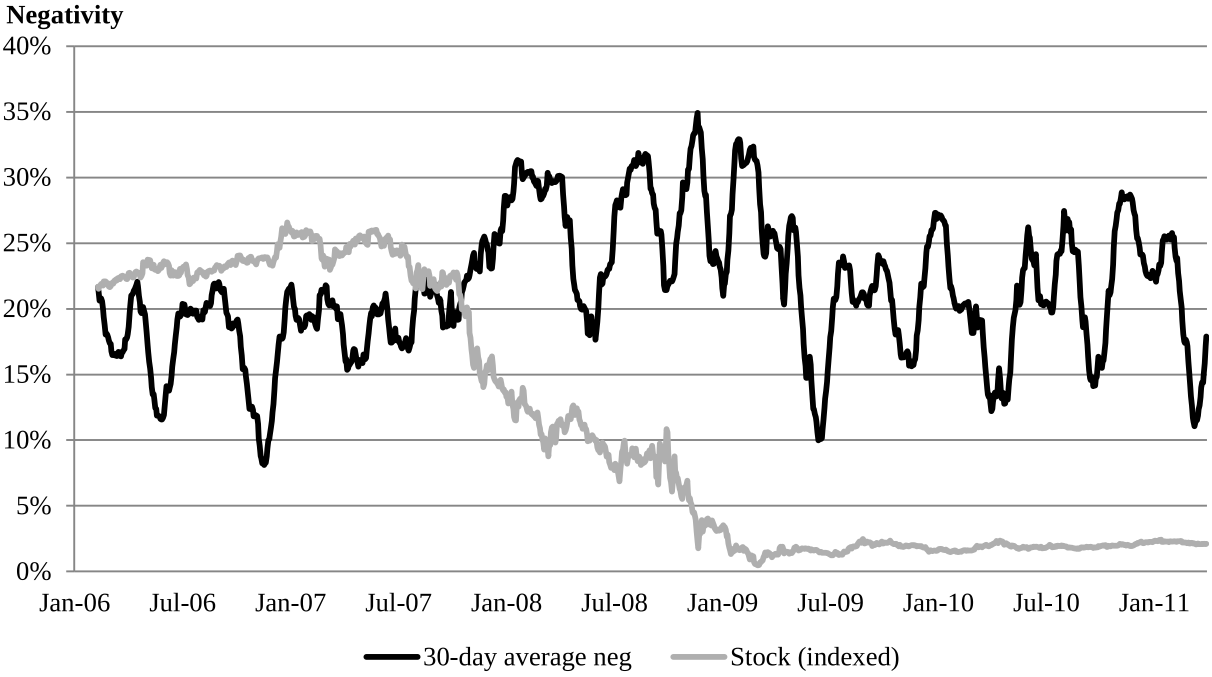


Figure 10: Citi sentiment and stock price

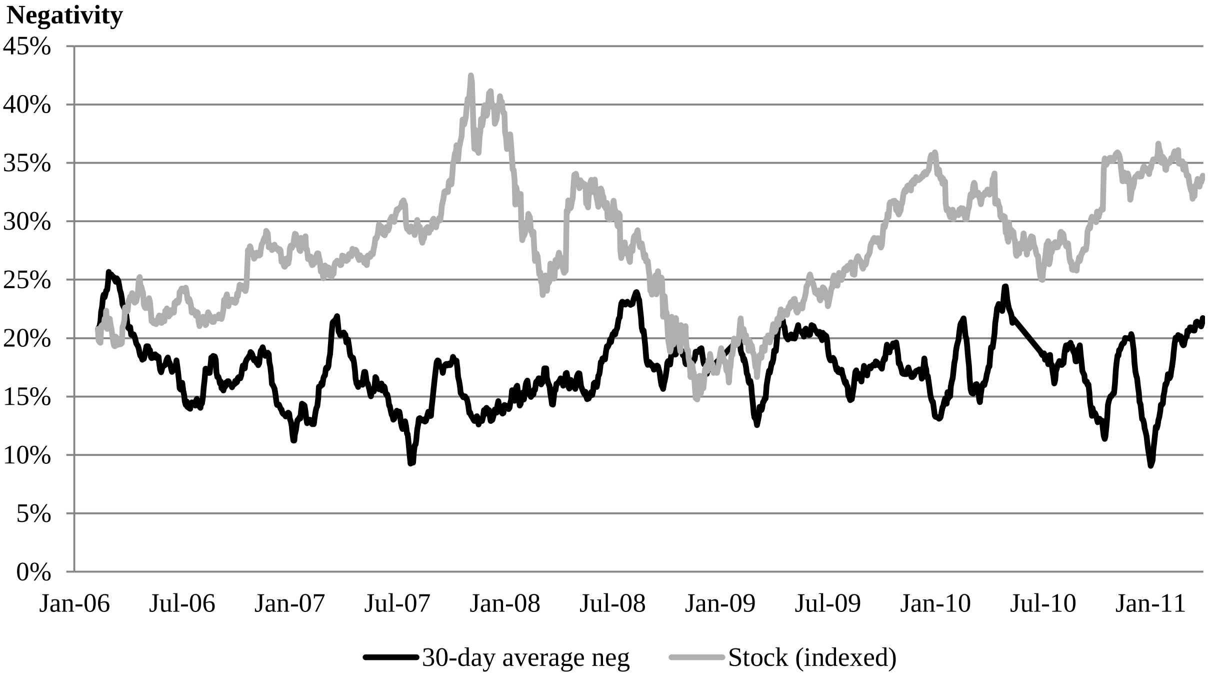


Figure 11: Google sentiment and stock price

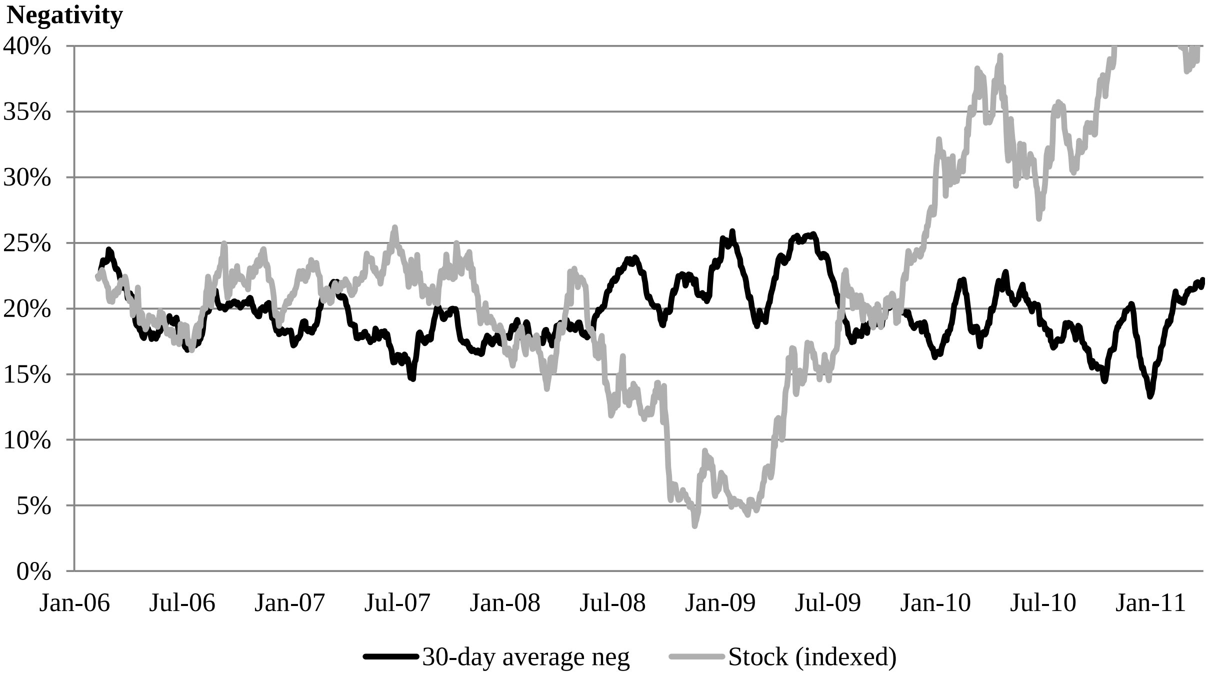


Figure 12: Ford sentiment and stock price

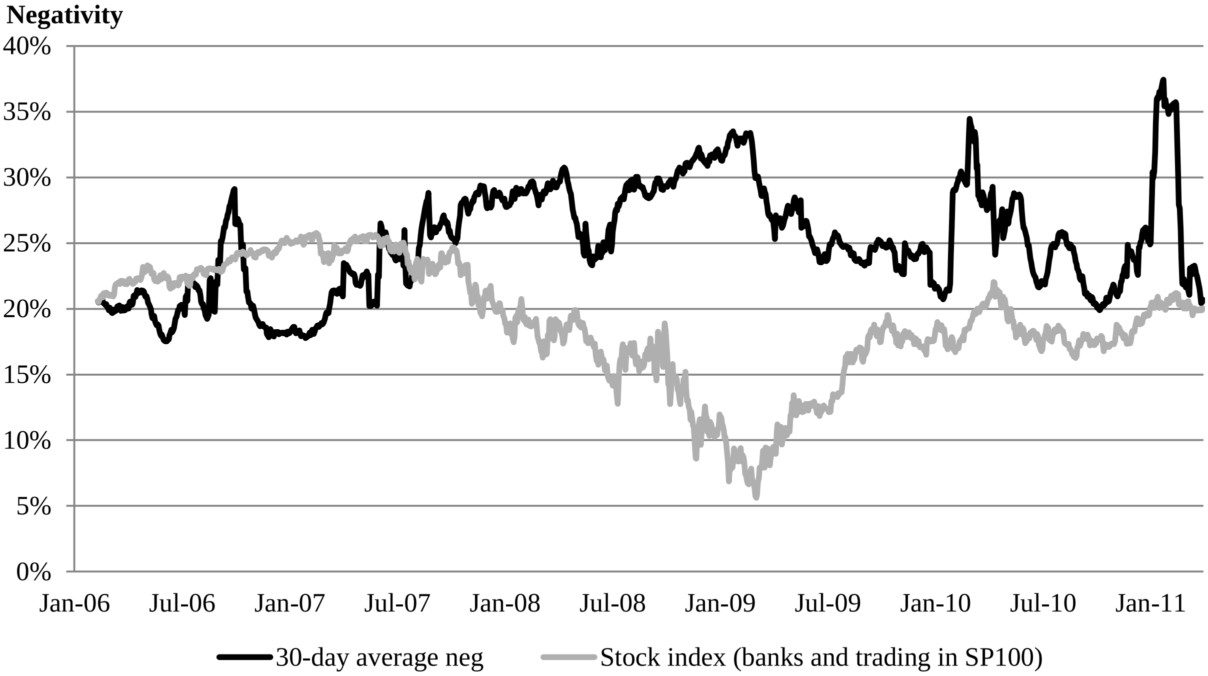


Figure 13: Banks and Trading sentiment and stock index

As can be seen with these examples, there appears to be some inverse relationship between negativity and stock price as one would expect. To further ensure the correctness of the sentiment score, we perform some further robustness checks.

To test whether our sentiment score could be driven by a bias in sources, we see how our sources differ in terms of sentiment. We study the difference between newspapers and other sources, and take as a case example our largest publication: The Financial Times. We plot negativity as calculated from each of these sources separately to see how well the sentiment scores in different sources correlate with each other.

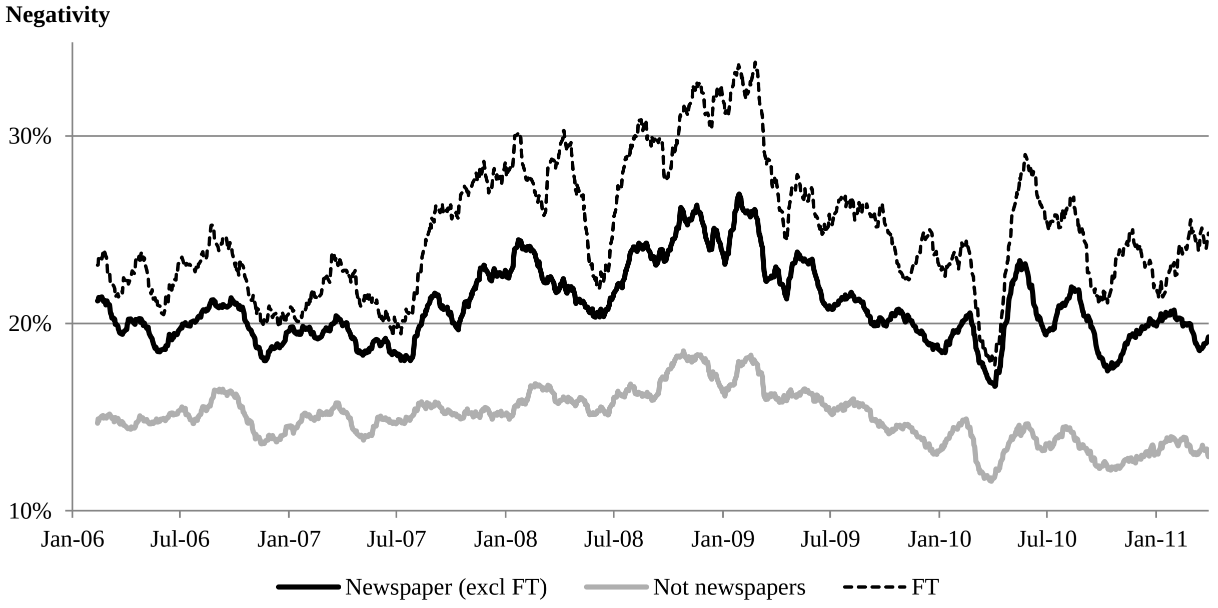


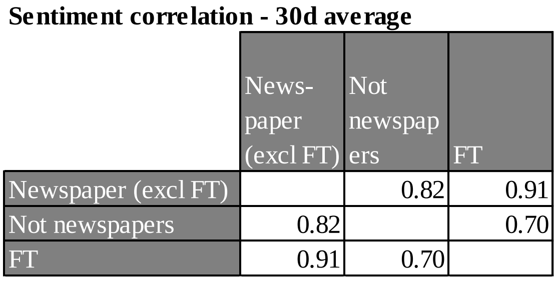
Figure 14: 30-day[154] average sentiment by different sources

As can be seen, all of the sources correlate with each other, reaching peaks and bottoms typically simultaneously. It is also noteworthy that newspapers: in particular The Financial Times, appear to be using significantly more negative language than other sources. We assume that the aforementioned result is impacted by companies' own earnings releases that typically would describe their operations in a more positive tone, and therefore impact the *'Not Newspapers'* category.

In addition, we calculate the correlation of these sentiments to verify the inference from the figure. We verify that the 30-day sentiments correlate relatively well with each other. Also as

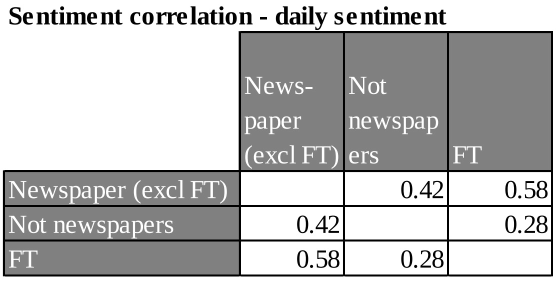
[154] The sentiment score varies significantly on a day-to-day basis. To show a more stable figure, a 30-day window is selected for most of our analysis. While a shorter window would have too much variation, a longer much longer window could already be impacted too much by multiple quarterly announcements etc.

we expect, if we test the correlation on a daily level, the sentiments correlate significantly less[155].



Sentiment correlation - 30d average

	News-paper (excl FT)	Not newspapers	FT
Newspaper (excl FT)		0.82	0.91
Not newspapers	0.82		0.70
FT	0.91	0.70	



Sentiment correlation - daily sentiment

	News-paper (excl FT)	Not newspapers	FT
Newspaper (excl FT)		0.42	0.58
Not newspapers	0.42		0.28
FT	0.58	0.28	

Figure 15: Correlation between different sentiments

5.2.4 Alternative ways for aggregating sentiment

In addition to sentiment aggregation described before, sentiment could be aggregated in alternative ways. For robustness, we test creating our sentiment scores with only a subset of the full media sample included in order to test if a certain media type could improve or bias our results. These variations of the sentiment score are described below. In general, these variations do not make a significant impact on our results. In situations where they do, we report also results of these variations.

Newspapers only and Only non-newspapers

We choose only sources that have been indicated to be newspapers. These sources are possibly more analytical compared to other news sources. Furthermore, this sentiment excludes all earnings announcements. On the other hand, we test the impact of choosing only the opposite: only media that is not from newspapers.

Sentiment only with at least 5 daily news

If there are less than five news for a company for a certain day, it is possible that a few articles analyzing the company from a certain perspective over a longer period of time could determine the sentiment. Often a reporter could e.g. spend a week writing about a company, and then publish an article. An article like this may not fully reflect the day’s sentiment anymore. To ensure that the news sentiment accurately pictures the overall media sentiment

[155] When comparing the sentiments on a company level each day, the correlations decrease naturally further as the number of publications each day for a particular company often varies between 0 to 10, i.e. the measure depends more on what publications are active on a particular day.

for a company on a certain date, we run some tests with the media sample so that we include sentiment scores only for days when at least 5 articles are available.

Windsorized results

To test if our results are driven by extreme values, we create a sentiment where we exclude the most extreme sentiment scores (windsorized at 1%).

30-day average sentiment

The sentiment score varies significantly on a day-to-day basis. To show a more stable figure, we test the sentiment of a 30-day average sentiment. While a shorter window would have too much variation, a longer window could already be impacted too much by multiple quarterly announcements etc.

Sentiment with FT news only

Our sample includes a significant portion of news from the Financial Times. It could be possible that a more credible publication would either picture the market more correctly or impact the market in the coming days. To test this, we run aggregate sentiment scores so that we include only Financial Times news to the score.

Sentiment only when low media disagreement

According to Das (2010), the market sentiment's predicting power may be weaker when there is diversity in opinions in the market. To measure the difference in opinions, Das proposes measuring a metric he calls '*Disag*' calculated as:

$$Disag \ = |1 - |\frac{Pos - Neg}{Pos + Neg}\ ||$$

where '*Pos'* is the number of positives: in our case, positive sentences, and '*Neg*' is the number of negatives: negative sentences. A zero value would indicate that all news would be in perfect agreement in terms of sentiment, while value of 1 for '*Disag*' would indicate that there are equally many positive and negative opinions. Based on our SVM's categorization of positive and negative sentences, we calculate *'Disag'* for each news item. Similarly to calculating negativity, we aggregate '*Disag*' for each company for each day by taking the average disagreement of all news. We also show the '*Disag*' figure for the whole SP100 in Figure 16, where we can see the market uncertainty during the financial crisis.

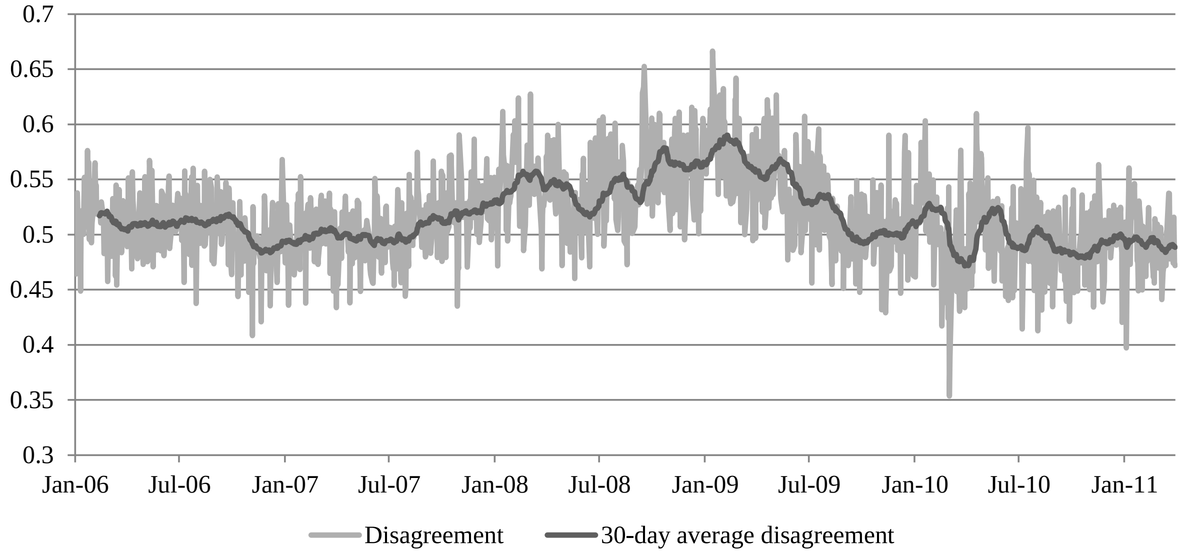


Figure 16: DISAG for SP100

SVM strong

To test the impact of having a highly robust sentiment metric, we create a measure “SVM strong”. For this, we combine the “Sentiment only when low news media disagreement” and “Sentiment only with at least 5 daily news”: we only include news where the disagreement is low (only lower 80% of sample included) and where there are at least 5 news items for the date.

5.3 Sentiment estimation methodology limitations

Das (2010) illustrate the inverse relationship between data volume and algorithm complexity in data and algorithm pyramid figure which is depicted below in Figure 17: The data and algorithms pyramids (Das, 2010). In general, we could categorize the *'bag-of-words*’ method as being on the lowest level of the pyramid, whereas our methodology: Linearized Phrase-Structure -model, would be in the content-level. However, as is evident from the figure, there is still work to be done to reach the context level.

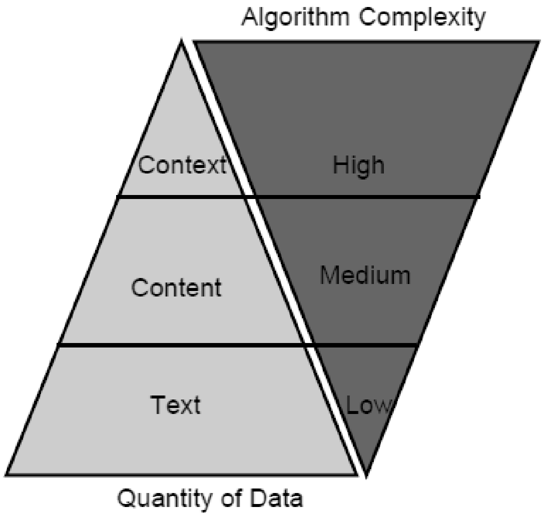


Figure 17: The data and algorithms pyramids (Das, 2010)

The prevalent methodology for the extant literature has been so far a naïve word count based on different dictionaries. While being simple and fast to use, the word count method has its limits. In their recent influential article, Loughran and McDonald (2011) show that dictionaries not related to the context of the data misclassify words. As a result, they create word lists for the financial context which significantly improve results for a word count methodology. Yet, the sentiment derived from a word count with context specific dictionaries remains a naïve proxy for the actual sentiment: simply counting words of a text cannot yield an understanding of the meaning of the text. Examples of the pitfalls of the methodology are multiple. For instance, word '*bad*' counts as a negative word in both the expressions '*bad result'* and *'not a bad result', or* sarcastically written text could be downright misinterpreted.

As an alternative way of measuring sentiment, we have proposed that Linearized Phrase-Structure -model can yield better results: recognizing common patterns in financial text that a word count is not able to do. While our methodology is an improvement vis-à-vis the prevalent methodology, it is still far from the actual sentiment that would be derived by multiple human annotators. Compared to a human annotator, the Linearized Phrase-Structure -model cannot detect topics, the relevance of a text, and is unable to assess text credibility. Also, we are unable pinpoint temporal differences in information in a text. All in all, our methodology is a significant improvement from the naïve word count methodology; however, there is yet significant room for improvement.

Our relevancy filtering is relatively limited, and we do not make a difference between topics and their relative importance. Ideally, we would retrieve all news that impact the sentiment of a company, and then sort them based on relevancy. At the moment, we filter implicitly as we search for news based on the companies' tickers. Should a news item be important but not mention the company: i.e. important industry news, we may miss the news item from our sentiment. Second, we give all news the same weight, regardless of their relative relevance. In reality, we might be better off by giving each news item a sentiment score, with a weight depending on the topic that it is written about. Possibly, we would have this kind of relevance assessment on two levels: first, we would assess what topics are relevant for the company, and by how much. Second, we could recognize on a sentence level how relevant each sentence is for the topic. Naturally, implementing topic recognition, and finding relative relevancy weights, would not be a trivial task, and could be a topic for future research.

The Linearized Phrase-Structure -model cannot assess credibility in sentences. Therefore, the algorithm operates in a child-like manner: believing everything that it sees. The aforesaid can lead into several biases that cause sentences to become tagged differently compared to that of a human annotator. Ideally, we would assess credibility of different authors, adjusting the opinions of authors that have a tendency to write in a certain way. For example, some publications may be more favorable in their writing style. For instance, a case-in-point is a situation where a company is the author of an article discussing its own operations in a positive manner.

Another caveat example relating to credibility is our aggregation method. Currently, we are assigning each article equal weight regardless of the publication; a naïve way of assigning weights to publications. For instance, an article in The Financial Times would most likely have more impact than a local newspaper article due to its larger circulation and higher perceived credibility. However, adjusting the methodology to take into account the aforementioned factors is an extensive undertaking; we leave the issues for future research. We suggest that, for example, different weights could be applied to the sentiment scores before aggregation, depending on the source.

Linearized Phrase-Structure -model does not make a difference between the temporal placements of information a news item is describing: we consider all found news with equal weight, regardless of the time period of the information in question. For instance, a story describing a company's history would be handled in the exact same way as a story bringing new information to the market, or a story speculating on the future. In reality, markets react very differently to new information vis-à-vis old information.[156] One approach to overcome the aforementioned limitation would be to detect topics as they appear for the first time, and discount secondary news: '*news of news'*. Another approach would be to keep track of the time aspect when estimating sentiment, and use that information in the estimation process. For example, Cahan et al. (2011) hypothesize that gathering speculations around future dates can help with the use of sentiment information.

In addition to the aforementioned considerations, our choice to focus on the fraction of negative sentences and words can be questioned.[157] It is a valid point that there may exist other metrics that would be more useful in the estimation of sentiment than negativity. However, extant literature has documented in several occasions that negativity outperforms other metrics (e.g., Tetlock, 2007). Therefore, we conclude that our choice to focus on the negativity of a given text is well-founded.

Another consideration is the qualitative data we use to estimate sentiment. As our sample consists of qualitative texts from LexisNexis database, we may miss some important qualitative text publications that are not included in the LexisNexis database. That being said, LexisNexis does cover different sources of qualitative texts quite comprehensively. Nevertheless, we may miss some publications due to copyright and coverage issues. Furthermore, we are missing qualitative texts that focus on specific products of companies but do not mention the company by its name. Such texts, and the sentiment in them, most likely carry significant value, and affect financial metrics. Also, we acknowledge the fact that we are missing the following qualitative sources completely from our sentiment: social media, non-written media: i.e., TV- and radio-broadcasts However, accounting for the aforesaid factors is not a trivial task, and therefore we suggest future research to study the matter.

We conclude that an ideal sentiment model would mimic the key stages of an analyst's thought process in assessing the impact that a news article has on a company, and would draw similar conclusions as a financial analyst would. In addition, such a model should incorporate

[156] However, we do recognize that tone can in itself have significant impact, even in the absence of new information (content).

[157] Previous studies have identified that negative sentiment appears to be the most influential one. However, for example, Das (2010) discovers that a daily 'disagreement-sentiment' that proxies the disagreement in the market by calculating the number of both positive and negative signals, can be used to estimate how well the negative sentiment works. In times when there are both positive and negative information on the market, the predictive power of the negative sentiment tends to decrease. We have aimed to take this into account by using the negative fraction of all words, but it is possible that by taking into account the positive sentiment in the way Das suggests, could further improve our results.

all available information that is relevant to a firm. We find that while our sentiment estimates are a significant improvement from word count methodology, more work remains to be done on estimating sentiment more accurately.

5.4 Relationship between sentiment and stock performance metrics

In order for us to examine the relationship between our main variables and the dependent variables in multivariate context, we need to choose a statistical research methodology. As the nature of our data is panel, we need to identify the appropriate method for calculating standard errors to avoid biased and unreliable statistical inferences as suggested by Petersen (2009). Therefore, we test our data to see what the correlations between the error term estimates are.

As suggested by Petersen (2009), we first estimate our standard errors[158] without clustering in any dimension. We proceed from there on to estimate the standard errors using clustering by firms: we find no evidence of magnitude change in our standard errors. Therefore, we conclude that our sample is free from firm effect (time-series correlation) — temporary or permanent. . We continue by clustering by day and quarter.[159] We find marginal increase in our standard errors for the daily clustering. However, we find a substantial increase in our standard errors for clustering by quarter. We continue the analysis by including calendar quarter dummies in our regressions while removing clustering. The increase in standard errors is mitigated. We conclude that our data is free from firm effect (time-series correlation; autocorrelation) but exhibits a permanent time effect (cross-sectional dependency) in error terms that is most salient in quarterly basis.[160] Hence, according to Petersen (2009), we rely on the most reliable method for estimating relationships under time effect: the Fama-Macbeth (1973) methodology. Therefore, we run our regression specifications under Fama-Macbeth methodology with clustering done in quarters. The quarters are specified as follows: Q1: *1st January to 31st March*; Q2: *1st April to 31st June*; Q3: *1st July to 31st September*; Q4: *1st*

[158] White (1984) adjusted standard errors to account for heteroskedasticity.

[159] Daily and quarterly clustering intervals have been used in majority of influential studies (e.g., Tetlock et al., 2008; Loughran and McDonald, 2011), and they are the most intuitive intervals considering the nature of our data: mainly daily or quarterly based.

[160] We reason that the quarterly vis-à-vis daily magnitude change is a function of relatively small number of observations per daily clusters: even as low as under 10 observations. Therefore, our coefficient estimates experience great noise when clustering with daily intervals. Hence, the estimates and their standard errors are not reliable. Therefore, we hypothesize that the firm-effect is indeed most salient on a daily basis but we are constrained by our sample. As a result, we rely on quarterly clustering.

October to 31st December. For our sample this means that we split the data into 21 time periods.

Our approach is similar to that of Loughran and McDonald (2011) with the exception that we do not discover a firm effect in our data and hence do not adjust our standard errors for autocorrelation using Newey-West (1987) approach.[161] In fact, as most of the content analysis papers have researched investor sentiment's impact on equity returns, most of the studies have not found firm effects in their data, as Petersen (2009) suggests when dealing with equity returns. Therefore, our finding of time effect in the data, and the corresponding choice of methodology, is consistent with the majority of recent papers that have relied on clustering by time to control for time effect (e.g., Engelberg, 2008; Tetlock et al., 2008; Hirsleifer et al., 2009).

We will continue this section by describing our study's multivariate main specifications and the alternative specifications designed to provide additional robustness to our results. All of the specifications have been conducted using the aforementioned Fama and MacBeth (1973) methodology with quarterly clustering of data. For more discussion on variable definitions, we refer the reader to Section 4.2.

5.4.1 Main specifications

We will discuss below the main specifications we are running to examine the relationship between the financial metrics of our choice (dependent variables), and the main independent variables. We will offer brief motivation for the included variables, as well as some reference literature with similar specifications in the domain of content analysis in finance.

For further information on the variable definitions, we refer the reader to section 4.2., and on information concerning the estimation of investor sentiment and the statistical methods used to study the relationship between financial metrics and the sentiment, we refer the reader to the previous sub-sections of this section.

Abnormal returns

Abnormal return (dependent variable) for the main specification is defined as a buy-and-hold abnormal return [BHAR] as discussed in Section 4.2. with 25 matching portfolios used as a

[161] Loughran and McDonald (2011) do not explicitly state that they had run an analysis on their data to discover the different forms of dependency in their error terms. However, we infer from their choice of methodology that they found both firm and time effects in their data.

benchmark return. The event windows are: [0,1], [1,5], [2,32] and [2,62], and the returns are calculated based on closing prices, as discussed in Section 4.2.

The main variables (independent variables) of the specification are as follows: first, we will attempt to estimate the impact of the prevailing investor sentiment by including a sentiment variable that will be based on two different methods: Linearized Phrase-Structure -model, and a bag-of-words word count using two separate dictionaries: the Harvard negative dictionary [H4N] and the Loughran and McDonald (2011) finance-negative dictionary; second, we will attempt to capture the distraction effect suggested by Hirsleifer et al. (2009) by including a market news volume count; third, we will include a firm specific news count to include the potential media coverage effect suggested by Fang and Peress (2009). The set of independent variables is designed to capture a holistic view of the impact of media on returns.

After the inclusion of our main variables, we include a set of controls to mitigate the risk of an omitted variable driving our results. Our list of controls for the main specification stands as follows: first, we include controls for size and book-to-market to capture the known risk factors as suggested by prior literature;[162] second, we include three different momentum factors to capture the effects of past returns on future returns; third, we include a control for share turnover to estimate liquidity and belief dispersion impacts on returns; fourth, we include a control for standardized earnings surprise to capture post-earnings announcement drift [PEAD]; fifth, we include a control for abnormal volatility to proxy for firm specific potential arbitrage limits; finally, we control for impact of institutions on returns with the hypothesis that institutions hold more information processing capacity and hence incorporate new information into prices more quickly.

The main specification we employ is similar in nature to that of Tetlock et al. (2008) and Loughran and McDonald (2011). The equation for our abnormal return specification is given below:

$$BHAR = c + \beta_1 Sentiment + \beta_2 Market\ News + \beta_3 Firm\ News + \beta_4 Controls$$

[162] We include these controls even though our matching portfolio is based on the same variables. However, by including size and B-2-M, we follow the approach of previous literature (e.g., Chan, 2003; Engelberg, 2008; Hirsleifer et al., 2009; Demers and Vega, 2010).

Abnormal volume

As trading volume is suggested to be related to investor sentiment (e.g., Tetlock, 2007; Hirsleifer et al., 2009; Loughran and McDonald, 2011), we design a specification to study the effects of our main variables on abnormal trading volume. As we do not foresee any specific reason as to why we should employ different event windows with volume vis-à-vis returns, we run our specification with the same four event windows as with returns in order to maintain consistency within the study.[163]

We define the dependent variable: abnormal volume, as explained in Section 4.2. The main variables for the specification are the same as in abnormal return specification. The set of controls is similar to that of returns with the following addition: in line with Hirsleifer et al. (2009), we include a control for abnormal market volume in order to isolate the idiosyncratic change in trading volume.

Our specification is similar especially to that of Loughran and McDonald (2011). The equation for our abnormal volume specification is given below:

$$Ab\ Vlm\ = c\ +\ \beta_1 Sentiment + \beta_2 Market\ News\ + \beta_3 Firm\ News + \beta_4 Controls$$

Abnormal volatility

As our final financial metric of interest, we study the relationship between our main variables and abnormal idiosyncratic volatility.[164] As suggested by Antweiler and Frank (2004), Demers and Vega (2010) and Loughran and McDonald (2011), qualitative text can have predictive power over future volatility of equity returns above and beyond quantitative information. We run abnormal volatility specification for event windows: [2,32] and [2,62], as discussed in Section 4.2.1. The shorter event windows are excluded as there are not enough data points for a meaningful estimate of volatility.

The set of main and control variables for the specification (independent variables) are the same as in abnormal returns. Therefore, our specification resembles mainly that of Loughran

[163] As previous section was dealing in the domain of equity returns, event windows were defined based on closing prices. Therefore, event windows [0,1] would read — in the domain of volume — as abnormal volume for day 1 [1]. With the exception of this slight change in notations, the event windows are the same. For more discussion, see Section 4.2.1.

[164] Refer to Section 4.2., for a definition of abnormal volatility

and McDonald’s (2011) specification. The equation for our abnormal volatility specification is as follows:

$$Ab\,Vol\ = c\ +\ \beta_1 Sentiment + \beta_2 Market\ News\ + \beta_3 Firm\ News + \beta_4 Controls$$

5.4.2 *Alternative specifications for robustness*

To ensure the reliability of our results, we employ several different specifications as robustness checks for our results. The goal of alternating specifications is to ensure that omitted variables are not driving our results, as well as to check whether or not our results dissipate when switching methodology in the estimation of our main variables. The section will continue by elaborating the different specifications we employ to ensure the robustness of our results. Reader should refer to Section 4.2., for more detailed descriptions on the variables discussed.

Alternative abnormal return definitions

Due to the critique towards the dependency of results on abnormal return calculation techniques (e.g., Fama, 1997, 1998), we run our abnormal return and volatility main specifications with differing return estimation techniques. First, we use raw returns (e.g., Tetlock, 2007) instead of abnormal returns; Second, we use two different benchmarks for abnormal returns: value weighted index returns based on all the market caps of the firms (e.g., Loughran and McDonald, 2011), and Fama and French three-factor model (e.g., Tetlock et al., 2008). Third, we estimate our abnormal returns using CARs instead of BHARs — we replicate this approach with the other aforementioned benchmarks as well.

The aforementioned alternations ensure that our results are not driven by a bad model problem. Due to the findings of prior literature, we expect to see changes in our results with different definitions of abnormal returns. At bare minimum, we expect a decrease in our coefficients magnitude when switching to CARs instead of BHARs.

Alternative main variable definitions

As we see in section 4.4.3, our news volume decreases slightly towards the end of the time period. To test whether or not the aforesaid impacts our tests, we run our main specifications with a standardized version of market and news volume variables. We standardize the volume by deducting the past average volume from the news volume, and dividing the resulting figure standardizing with standard deviation of volume. Therefore, we control seasonal trend with standardization, and examine the abnormal impact of volume to our results. We define standardized market and firm specific news volume on day 0 as follows:

$$Standardized\ volume_0 = \frac{\left(News\ Volume_0 - \ Average\ Volume_{-252,-2}\right)}{Standard\ Deviation\ of\ Volume_{-252,-2}}$$

Additional control variables

In order to be robust, we create several different alternative specifications that include new controls in addition to our main controls. Our aim is to examine whether or not some of the new controls have an impact on our results. However, as prior studies have not found significant evidence of changes in results with the inclusion of the additional controls we plan on including, we do not expect them to alter our main specification results.

- **Industry specific effects.** We add an industry dummy: a dummy according to the Fama and French (1997) 48-industry classification to our main specifications in order to account for industry specific differences in results. Due to the large number of dummies included, we will lose several degrees of freedom, and hence expect our results to decrease in statistical significance. However, qualitatively our results should remain the same; in other words, we do not believe that a specific industry would be driving our results.
- **Impact of analysts.** We include the number of analysts following a company, and the analyst dispersion control variables into our main specifications to see whether or not analysts have an impact on our dependent variables. However, our main specification should already account for the effects that analyst coverage[165] and analyst dispersion[166] cover. Therefore, we do not expect to see significant changes in our results.
- **Calendar effects.** We employ calendar dummies to our main specifications to be absolutely sure that our results are not dependent on the well-documented calendar effects. However, as our matching portfolio benchmark return should already cover calendar effects implicitly, we do not expect changes with abnormal return and

[165] Analyst coverage is a proxy for informational efficiency that should be covered by institutional ownership. Due to strong multicollinearity issues analyst coverage is not included in the main specification.

[166] Analyst dispersion is a proxy for belief dispersion. Share turnover should also proxy for belief dispersion.

volatility specifications. Moreover, as we control for market wide abnormal trading in our abnormal volume specification,[167] we do not expect to see changes in the volume specification, either.

- **Dividend effects.** In line with Li (2006), we include dividend variables as suggested by Fama and French (2006) in order to cover their possible impact on our dependent variables. Hence, we include the paid dividends from the last twelve months, divided by book value of equity. Also, we add a dummy for companies with no last twelve months dividends dummy. We do not expect to see a change in our results with the addition of dividend variables.

[167] Abnormal market volume should implicitly cover calendar effects.

6 RESULTS

In this chapter, we discuss the results we derive from our tests, and how they relate to our hypotheses. We begin by testing the performance of different sentiment methodologies in classifying articles. Afterwards, we move on to study the impact of sentiment on financial metrics. To do so, we begin by conducting univariate tests to determine if there is a relationship, and what is the nature of the relationship. After the univariate tests, we move on to perform multivariate tests to see how our main independent variables impact our dependent variables when tested simultaneously in a holistic media model. Finally, we conduct several additional multivariate tests with alternative specifications for the sake of robustness, and to further study specific areas of interest.

We refer the reader to Section 3 for more information on our hypotheses, to Section 4 on descriptions of data and our variables, and to Section 5 for discussion on our methodology and specifications.

6.1 Sentence-level sentiment

In order to compare the accuracy of the LPS algorithm, a number of baseline models were constructed. One of the objectives in the experiments is to understand how the added layers of rules contribute to the overall model performance. Therefore, the benchmark algorithms featured below have been chosen to represent different levels of model complexity ranging from simple word based algorithms towards the model proposed by Moilanen et al. (2010), and our approach.

We start by comparing different methodologies against the benchmark of annotated sentences. From there on we move to discuss the potential sources of errors in our method: Linearized Phrase-Structure -model.

6.1.1 Comparison of methods

We run several different sentiment classification methodologies for both our annotated datasets: Training set A and Training set B. We test the following ways of estimating sentiment:

- **Random**: To be able to compare the improvements of different methods over random assigning of labels, we calculate sentiment scores for randomly labeled sentences.

Using a uniform distribution, we assign a random sentiment for each sentence: positive, neutral or negative. We simulate the random assigning 1,000 times, and compare the outcome: i.e., accuracy, between the random sample and our methodologies.

- ❖ **Weighted random**: We use the distribution of data labels in the annotated data sets: i.e., if there are 10% of negative sentences in a training set, each sentence will have a 10% probability of being tagged as negative. Similarly to Random, we give each sentence a random sentiment, and we run simulation 1,000 times.
- ❖ **Word count (Harvard):** As explained in previous sections, we calculate the number of positive and negative words in each sentence using the Harvard dictionary positive and negative word lists. We then label the sentence using the following rules: no polarized words or ambiguous = neutral; 2/3 or more negative words = negative; 2/3 or more positive words = positive.
- ❖ **Word count (Loughran)**: We use a similar method as with word count Harvard, but use the positive and negative word lists by Loughran and McDonald (2011).
- ❖ **MPQA**: As the primary baseline in the experiments regarding sentiment classification accuracy, we consider the Quasi-compositional Polarity Sequencing model with MPQA dictionary proposed by Moilanen et al. (2010). In the paper, they compared a number of alternative models with varying levels of complexity, but taken as a whole they found that a simple polarity-sequence model outperformed their more complicated models relying on complete phrase-structure information. The version considered here is this model with MPQA dictionary as the source of polarity information. When evaluating performance, we always use 90% of the training material and run the algorithm on the remaining 10%, not to test the algorithm on the data that it was trained on (10-fold cross-validation).
- ❖ **LPS**. Finally, we run our Linearized Phrase-Structure -model, described in section 5.1.2, with all word lists: lists by Loughran and McDonald, augmented with the additional wordlists we described in section 5.1.3. As above, we always test the algorithm on 10% of training material (10-fold cross-validation).
- ❖ **True values**. For easy comparison, we also show all metrics for sentences that have been labeled with the correct labels.

For all the methods above, we calculate a set of accuracy metrics to show how well the methods perform on recognizing positive, neutral and negative sentences. The outcomes of the metrics are defined as follows: ‘*positive’* refers to a sentence belonging to the category at hand; ‘*negative’* refers to a sentence not belonging to a category. For example, when we look at accuracy metrics of neutral sentences, a ‘*true positive’* would be a sentence that is neutral and has been labeled as neutral, and a ‘*false positive*’ would be a sentence that is not neutral (i.e. is positive or negative) but has been labeled as neutral, etc. We use the classical confusion matrix for assessing classification accuracy as seen in Table 12: Confusion matrix (see also: Das, 2010). In the matrix, all cells on the diagonal are considered as being correct.

Table 12: Confusion matrix

Confusion matrix according to Fawcett, 2005

		Annotated Class	
		Positive	**Negative**
Estimated Class	**Positive**	True Positives	**False Positives**
	Negative	**False Negatives**	True Negatives

From the aforementioned, we calculate the following metrics:

- **Accuracy**: Estimates what fraction of items in a category has been correctly identified as belonging to a category, or not belonging to a category. For example, if calculated for the group of neutral sentences, this would refer to correctly identified neutral sentences, and correctly identified non-neutral sentences, as a fraction of all sentences. In general, accuracy of an intelligent algorithm should be above (1 / number of classes), the accuracy of random guessing (Das, 2010).

$$\frac{(True\ Positives + True\ Negatives)}{All\ positives\ and\ negatives\ (true\ pos. + true\ neg. + false\ pos. + false\ neg.)}$$

- **Recall**: Estimates what fraction of items in a category have been correctly identified. For example: correctly identified positive sentences / all positive sentences in the sample. A figure below 1 would mean that some figures have been classified falsely (e.g. a neutral sentence has been classified as a positive sentence).

$$\frac{True\ Positives}{All\ positives\ (true\ positives + false\ negatives)}$$

- **Precision:** Estimates what fraction of items in a category have been correctly identified vs. all items that have been identified in the category: e.g. correctly identified positive sentences / all identified positive sentences. For example, for neutral sentences, a figure below 1 would in this case mean that some of the neutral sentences have not been recognized as being neutral.

$$\frac{True\ Positives}{Identified\ positives\ (true\ \ positives + false\ positives)}$$

- **F1-score:** The F1-score is a measure that combines recall and precision.

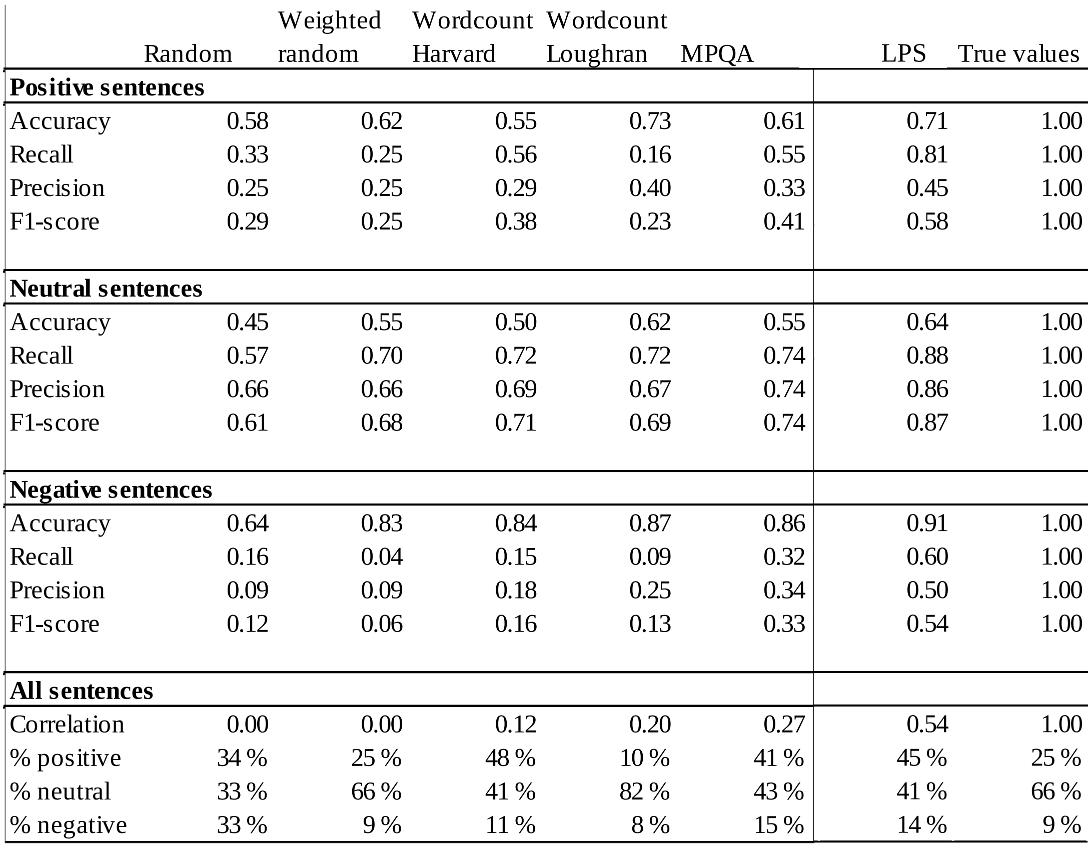
.

$$\frac{2 \times True\ Positives}{\left(Positives + (True\ Positives + False\ Positives)\right)}$$

Furthermore, we calculate for all samples the fraction of different categories: fraction of positive, neutral and negative labels given, and the correlation between the assigned figure and the true figure[168]. The results reported for the algorithms with a machine learning component (i.e. MPQA and LPS) are computed using 10-fold cross-validation. The results of the methods can be found in Table 13: Performance of different method - Training set A

[168] In order to calculate a numerical correlation figure, we simply transform the definitions to numbers as follows: '*positive*' = 3, '*neutral*' = 2, and '*negative*' = 1.

Table 13: Performance of different method - Training set A



		Weighted	Wordcount	Wordcount			
	Random	random	Harvard	Loughran	MPQA	LPS	True values
Positive sentences							
Accuracy	0.58	0.62	0.55	0.73	0.61	0.71	1.00
Recall	0.33	0.25	0.56	0.16	0.55	0.81	1.00
Precision	0.25	0.25	0.29	0.40	0.33	0.45	1.00
F1-score	0.29	0.25	0.38	0.23	0.41	0.58	1.00
Neutral sentences							
Accuracy	0.45	0.55	0.50	0.62	0.55	0.64	1.00
Recall	0.57	0.70	0.72	0.72	0.74	0.88	1.00
Precision	0.66	0.66	0.69	0.67	0.74	0.86	1.00
F1-score	0.61	0.68	0.71	0.69	0.74	0.87	1.00
Negative sentences							
Accuracy	0.64	0.83	0.84	0.87	0.86	0.91	1.00
Recall	0.16	0.04	0.15	0.09	0.32	0.60	1.00
Precision	0.09	0.09	0.18	0.25	0.34	0.50	1.00
F1-score	0.12	0.06	0.16	0.13	0.33	0.54	1.00
All sentences							
Correlation	0.00	0.00	0.12	0.20	0.27	0.54	1.00
% positive	34 %	25 %	48 %	10 %	41 %	45 %	25 %
% neutral	33 %	66 %	41 %	82 %	43 %	41 %	66 %
% negative	33 %	9 %	11 %	8 %	15 %	14 %	9 %

In addition, we calculate the same metrics for Training set B. These results are shown in Table 14.

Table 14: Performance of different methods - Training set B

	Random	Weighted random	Wordcount Harvard	Wordcount Loughran	MPQA	LPS	True values
Positive sentences							
Accuracy	0.54	0.53	0.58	0.68	0.68	0.85	1.00
Recall	0.33	0.38	0.59	0.21	0.62	0.89	1.00
Precision	0.38	0.38	0.46	0.79	0.56	0.75	1.00
F1-score	0.35	0.38	0.52	0.34	0.59	0.81	1.00
Neutral sentences							
Accuracy	0.51	0.50	0.55	0.57	0.64	0.82	1.00
Recall	0.39	0.50	0.59	0.60	0.66	0.90	1.00
Precision	0.48	0.48	0.53	0.53	0.64	0.87	1.00
F1-score	0.43	0.49	0.56	0.56	0.65	0.88	1.00
Negative sentences							
Accuracy	0.62	0.75	0.81	0.86	0.85	0.95	1.00
Recall	0.16	0.08	0.18	0.13	0.35	0.74	1.00
Precision	0.14	0.14	0.31	0.53	0.49	0.82	1.00
F1-score	0.15	0.10	0.23	0.21	0.41	0.78	1.00
All sentences							
Correlation	0.00	0.00	0.20	0.36	0.41	0.76	1.00
% positive	33 %	38 %	48 %	10 %	41 %	45 %	38 %
% neutral	33 %	48 %	41 %	82 %	43 %	41 %	48 %
% negative	33 %	14 %	11 %	8 %	15 %	14 %	14 %

As expected, our results improve in general from left to the right. We can see that all metrics estimating sentiment are superior compared to random guessing. Loughran and McDonald wordlists also improve results over the word count with Harvard dictionary, even though the difference is not a large improvement.

When comparing the results, it is clear that the SVM-based methods are more accurate ways to measure sentiment over word counts. The simplest SVM rules should effectively mimic the use of word count: e.g. a negative label will be assigned to sentences where the only identified polarized words are negative. Thus, it is not surprising that SVM performs at the same level as word count. As SVMs can detect also more advanced patterns in sentences, it is expected that SVMs perform better than word count. Also, we have been able to achieve significant improvement in performance by developing our SVM method from Quasi-compositional Polarity Sequencing to Linearized Phrase-Structure -model. These results are also in line with our expectations.

Comparing results of the two training sets, we also find that the Training set B performs significantly better. This is influenced especially by the fact that with this training set sentences that are irrelevant but have a sentiment are still classified by the human annotator as having a sentiment. As the LPS method does not have any feature that determines relevance of a certain sentence, it cannot make a distinction between two sentences with same polarities. Thus it would classify two sentences with the same polarity structures equally, even if the other one was discussing something irrelevant. For a further discussion on errors performed by the algorithm, we continue with an analysis of errors in the following section.

Human annotators do not always agree on how a sentence should be annotated. Even in instances where we give clear instructions to an annotator, the resulting accuracy of a second annotator, compared to the first annotator, is between 80-100%. The accuracy is impacted by: the sample complexity, the annotators' background, their interpretation of different events, etc. As can be seen from the results, Linearized Phrase-Structure -model is close to annotator-achieved accuracy. While some efforts can still improve the algorithm's performance in cases where annotators agree, the chosen principles used to select a certain label can be as important when developing a more accurate sentiment – both with the algorithm and between annotators. Ultimately, this leads to the question of how should sentiment be aggregated and what would be the most beneficial interpretation of it. Options range from reporting the sentiment as the '*feeling*' of the market, seeing the sentiment score as a force impacting the thinking of investors, trying to catch especially the sentiment around future events and comparing these expectations to realized performance, and so on.

To conclude, use of better lexicons alone is sufficient to boost performance considerably, as indicated by the use of Loughran and McDonald's dictionary instead of Harvard's dictionary. However, we find the role of good learning algorithms at least equally important. Already the MPQA-baseline is substantially stronger than either of the lexicon-based alternatives. Finally, we highlight the importance of choosing the right principles for identifying sentiment for different purposes.

6.1.2 Sources of error for Linearized Phrase-Structure -model

For the sentences that the algorithm annotates incorrectly, we take a random sample of 300 sentences, and ask our annotator to classify them based on the type of error that the algorithm is making. Key reasons for errors appear to be: the lack of knowledge on what is relevant for a company's success,[169] the inability to assess the credibility in a sentence,[170] and the failure to identify what is new information.[171] The results of our error analysis can be seen in Table 15: Error statistics for Linearized Phrase-Structure -model (Training set A % of errors refers to how many sentences were misclassified because of this reason. Often it is possible that two errors are simultaneously present in a sentence. Therefore, we also show how often the error type was the only error present.

Table 15: Error statistics for Linearized Phrase-Structure -model (Training set A)

Error-type[172]	**% of errors**	**% as the only error**
Need for more context	43.0 %	20.0 %
Inability to recognize significance of events	26.7 %	19.3 %
Company talking in advertising like -tone about its' own operations	10.0 %	3.3 %
Recognizing patterns in descriptive text	9.0 %	7.7 %
Polysemy of words and expressions	8.7 %	0.7 %
Positive convention of talking about something	6.3 %	2.0 %
Words lacking from word lists	6.0 %	1.0 %
Inability to understand magnitude and value of items	5.3 %	0.3 %
Inability to detect changes in numbers	4.3 %	3.7 %
Inability to detect roles in sentence	4.3 %	0.3 %
Wrong computer patterns	4.0 %	0.3 %
Use of longer expressions and interpreting non-words as words	4.0 %	0.0 %
Inability to detect time expressions in the sentence	3.3 %	2.7 %
Borderline cases: sentences that could be tagged by human as either/or	2.7 %	1.0 %
Sentences with multiple parts	1.7 %	0.3 %
Inability to reason from text	1.3 %	0.7 %
Wrong label in training data	1.3 %	0.3 %

In addition to the above error analysis, we review the sentences with errors in order to identify common individual words. In particular, we look for words that exist in our word lists, and

[169] The errors we are referring to are: 'inability to recognize significance of events' and 'need for more context.'
[170] The errors we are referring to are: 'company talking in advertising like -tone about its' own operations', 'positive convention of talking about something.'
[171] The errors we are referring to are: 'inability to detect time expressions in the sentence.'
[172] For a description of error types, refer to Appendix J - Error descriptions for LPS.

appear to lead to more errors compared to the sentiments that they help to identify. For a list of these words, see Appendix I – Wordlist defects. We refine the algorithm based on the aforementioned by removing the words from the wordlists we use to identify sentiment, and hence improve the accuracy of all of our methods that are wordlist-based[173].

6.2 Univariate tests

As a first step in studying the relationship between our dependent variables and our main independent variables, we turn to univariate tests. In order to understand whether or not our main independent variables have a relationship with our dependent variables, we sort our main independent variables into ten deciles, and calculate the median dependent variable value for all the deciles. After the sorts and calculations, we plot the deciles and their corresponding dependent variable median values on scatter plots to see the nature of the relationship. In the case that a clear relationship exists, the dependent variable values should move monotonically across the different deciles: i.e., with abnormal returns, returns should decrease monotonically as we move from less negative deciles towards more negative deciles.

We begin by looking at the relationship between abnormal returns and our main independent variables. From there on, we move on to discuss the relationship between abnormal trading volume and our main independent variables. Finally, we study the relationship of abnormal volatility and our main independent variables.

6.2.1 Abnormal returns

We begin our univariate study by looking at the relationship between abnormal returns and our main independent variables. We will begin by plotting our primary sentiment estimation methodology: Linearized Phrase-Structure -model, deciles against decile median abnormal return values. The graphs are shown below in Figure 18. After looking at Linearized Phrase-Structure -model performance, we will look at how our two other main independent variables fare in the univariate tests by plotting market news volume (Figure 19) and firm specific news volume (Figure 20) against abnormal returns. The graphs are shown below.

[173] The aforesaid modification is reflected in all shown results.

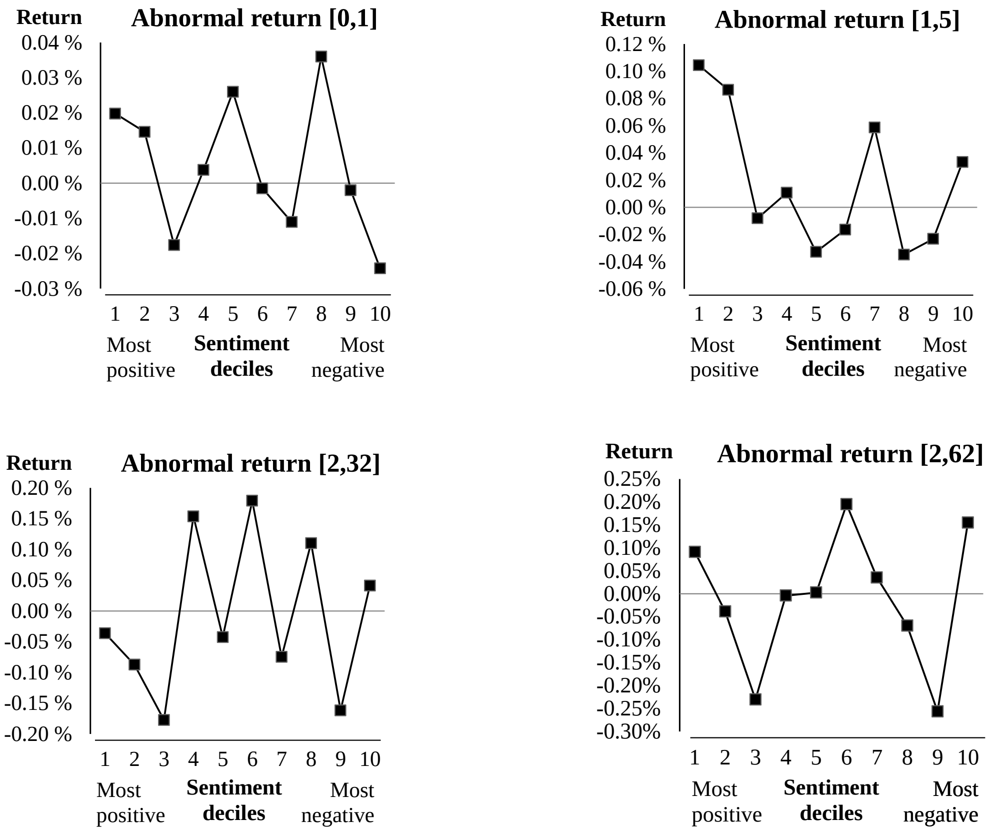


Figure 18: LPS model deciles vs. decile median abnormal return

As is clear from Figure 18, sentiment does not seem to have a relationship with abnormal returns in any of the event windows. The closest to a monotonic relationship is event window [1,5]. However, even with the 4-day event window, the relationship is far from monotonic, and decile 7 values are off the chart. We can infer from the univariate tests that sentiment does not seem to explain future variations in abnormal returns. Therefore, our univariate tests are in contradiction with prior literature’s findings, and seem to support market efficiency. We conclude that we need to move on to multivariate tests to specify the exact nature of the relationship in order to draw more precise conclusions. However, based on the univariate tests, it seems that our findings are supporting the hypothesis that markets are efficient. In other words, sentiment cannot be used to forecast abnormal returns.

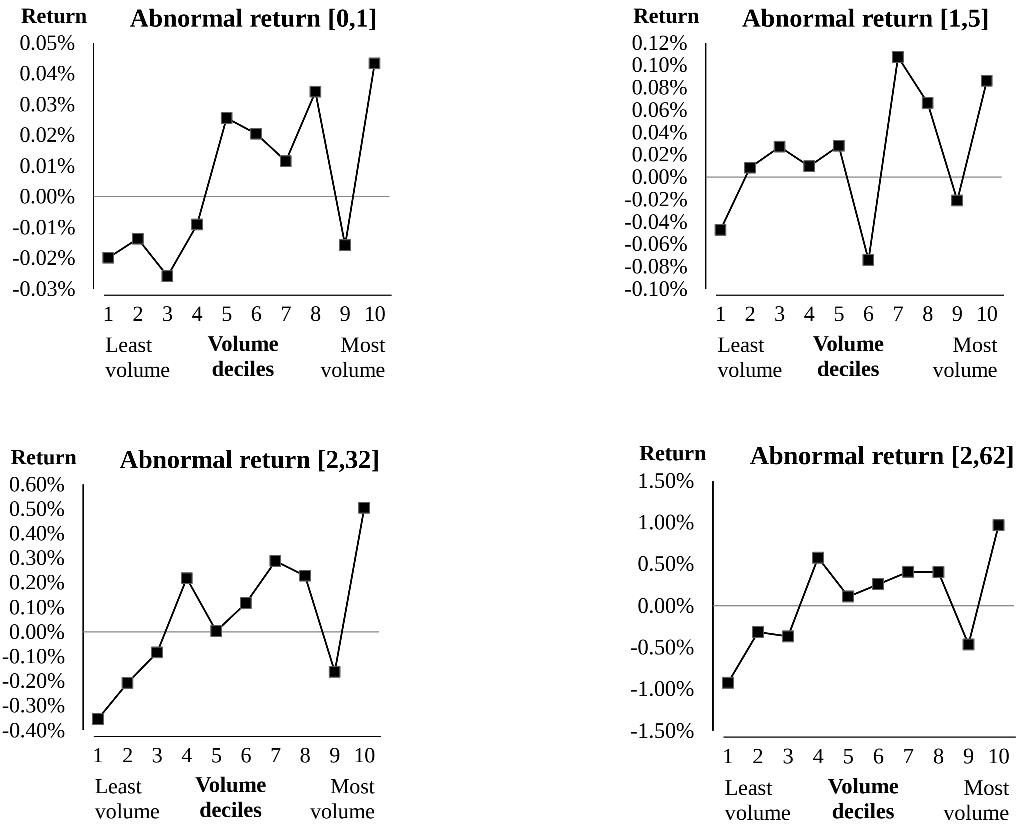


Figure 19: Market news volume deciles vs. decile median abnormal return

Market news volume holds more promise than sentiment variable in forecasting abnormal returns based on the univariate tests. However, decile 9 exhibits a peculiar drop in all event windows. Studying further data behind this decile, we notice an interesting pattern: a much larger fraction of news from 2009 and 2010 are present in the the lowest deciles and the 9th decile. As these have been the core years of the financial crisis, it is natural that these returns are lower than for the other deciles. We hypothesize that during this time period on days when there has been a large coverage this has often been due to negative press. Indeed this appears to be the case. We next drill down in to returns below -10% per decile. The 9th decile indeed has proportionately larger negative returns than the other top deciles. We conclude that the drop in the 9th decile is thus very likely caused by the fact that our sample includes the time period with the financial crisis. Nevertheless, the relationship between market news volume and abnormal returns resembles a monotonic relationship to much greater degree than the pattern with sentiment variable. Furthermore, it seems that the relationship is positive, as we hypothesized. Therefore, univariate tests seem to support our hypothesis of underreaction with aggregate market news volume.

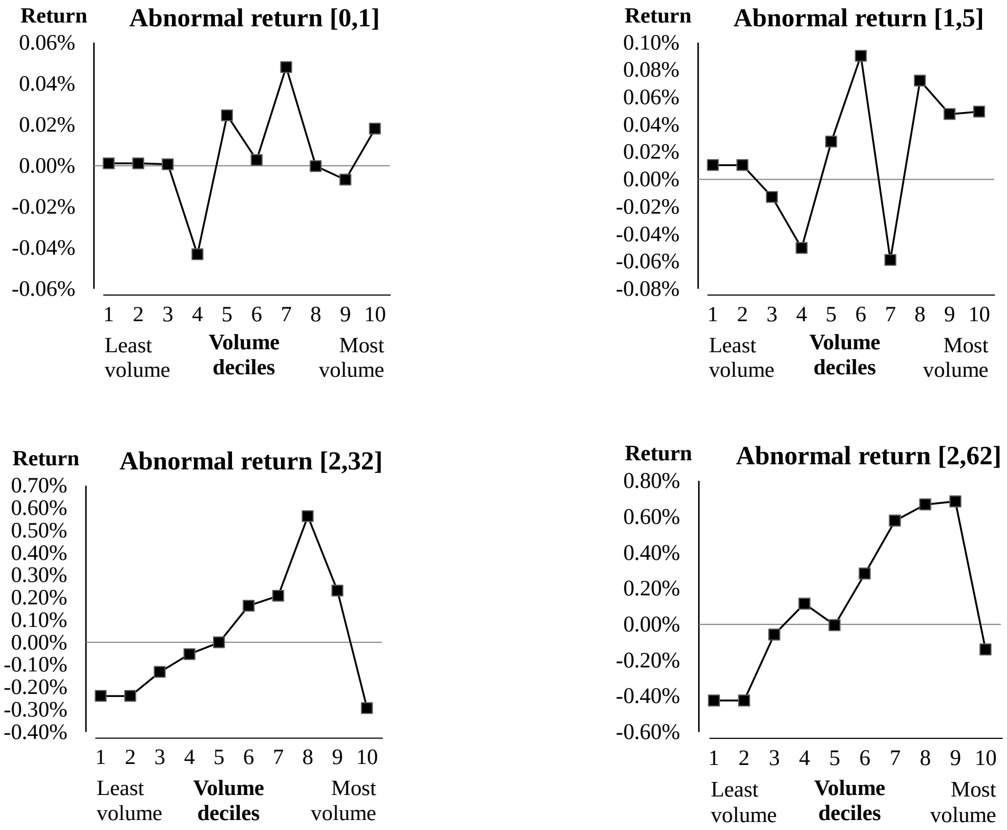


Figure 20: Firm specific news[174] volume deciles vs. decile median abnormal return

Firm specific news volume shows no clear relationship with the shorter event-windows. However, the longer event windows seem to exhibit an increasing relationship, yet the radical drop in the last deciles is puzzling. Looking further into the data, we notice that the highest news volume is concentrated especially to certain companies, including also several financial institutions where returns have been significantly more negative than for others. Thus, this drop appears to represent a bias in our sample during the financial crisis. All in all, the relationship between firm specific news volume and abnormal returns is unclear based on the univariate tests. We anticipate that firm specific news volume will not have a significant relationship with abnormal returns on the short-term event windows based on our univariate findings. Also, we remain skeptical towards a significant relationship in the longer event windows.[175]

[174] We use only data with news volume over 5, as otherwise too many deciles would have equal values. In other words, the first 4 deciles would have news volume value of 0, and the next two deciles would have a news volume value of 1.

[175] Our findings with firm specific news volume and abnormal trading volume push us to hypothesize on the nature of the relationship that firm specific news volume has with dependent variables. We refer the reader to the section dealing with abnormal trading volume and firm specific news volume for this discussion.

6.2.2 Abnormal volume

We continue our univariate tests by looking at the relationship between abnormal trading volume and our main independent variables. We will proceed in a similar manner as in the previous section: first, we plot Linearized Phrase-Structure -model sentiment against abnormal volumes; second, we plot market news volume against abnormal volumes; finally, we plot firm specific news volume against abnormal volume. The graphs are shown below.

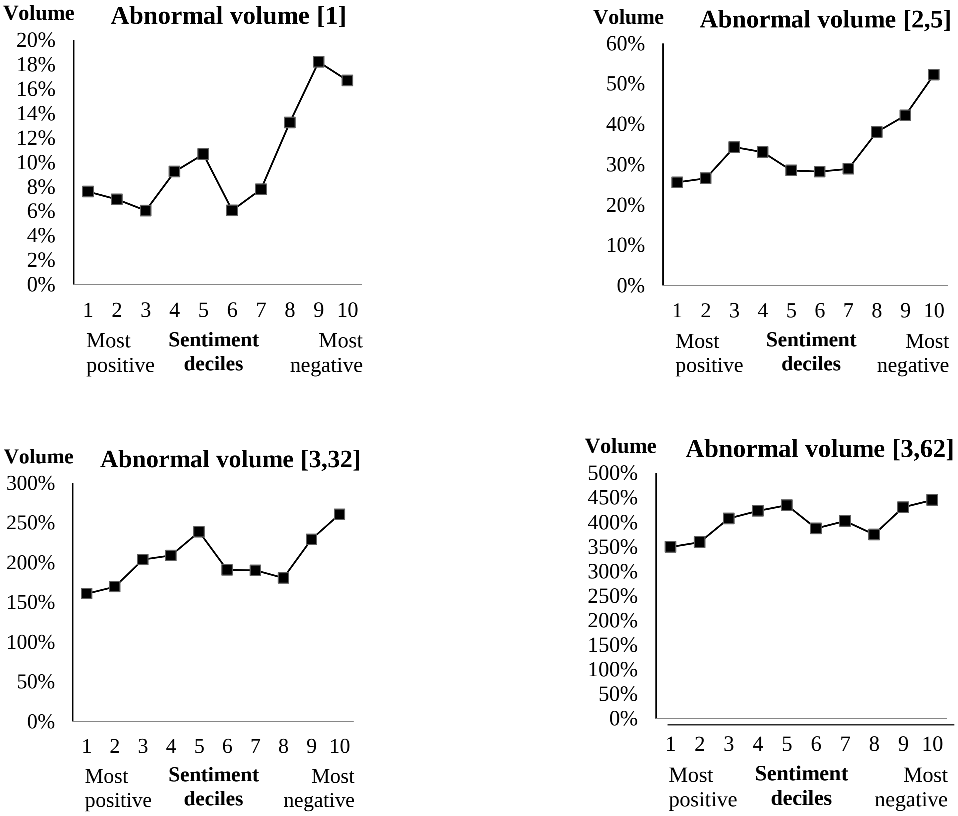


Figure 21: LPS model deciles vs. decile median abnormal trading volume

Based on Figure 20, it seems that sentiment might have a relationship with abnormal volume. From the graphs we can see that the relationship is not perfectly monotonic but does show a clear increasing pattern - especially event window [2,5] seems to fare well. Based on the findings, we infer that our hypothesis of information content dominating investor reactions over tone seems to be correct, and sentiment has a positive relationship with abnormal volume.

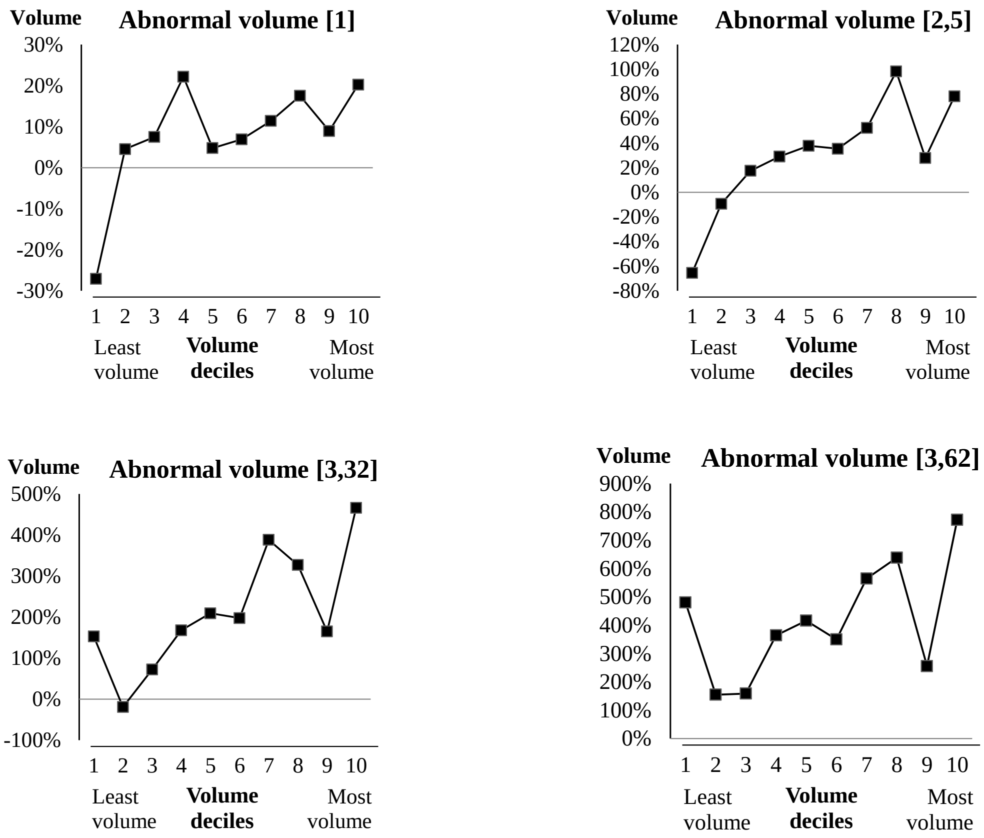


Figure 22: Market news volume deciles vs. decile median abnormal trading volume

Market news volume seems to exhibit a relationship with abnormal volume as well. However, the relationship portrayed by sentiment deciles seems to hold more promise than market news volume. There is considerable variation in decile median values in the event windows. Again, decile 9 seems to experience a significant drop that disrupts the increasing monotonic trend[176]. Nevertheless, an increasing pattern does emerge from the figure, and we infer that the univariate tests do indicate some level of support for our hypothesis of an underreaction, followed by abnormal volume in the longer time frame.

[176] As for our results comparing Market news volume and to abnormal returns, we suggest that this jump is likely caused by the fact that our sample includes the financial crisis.

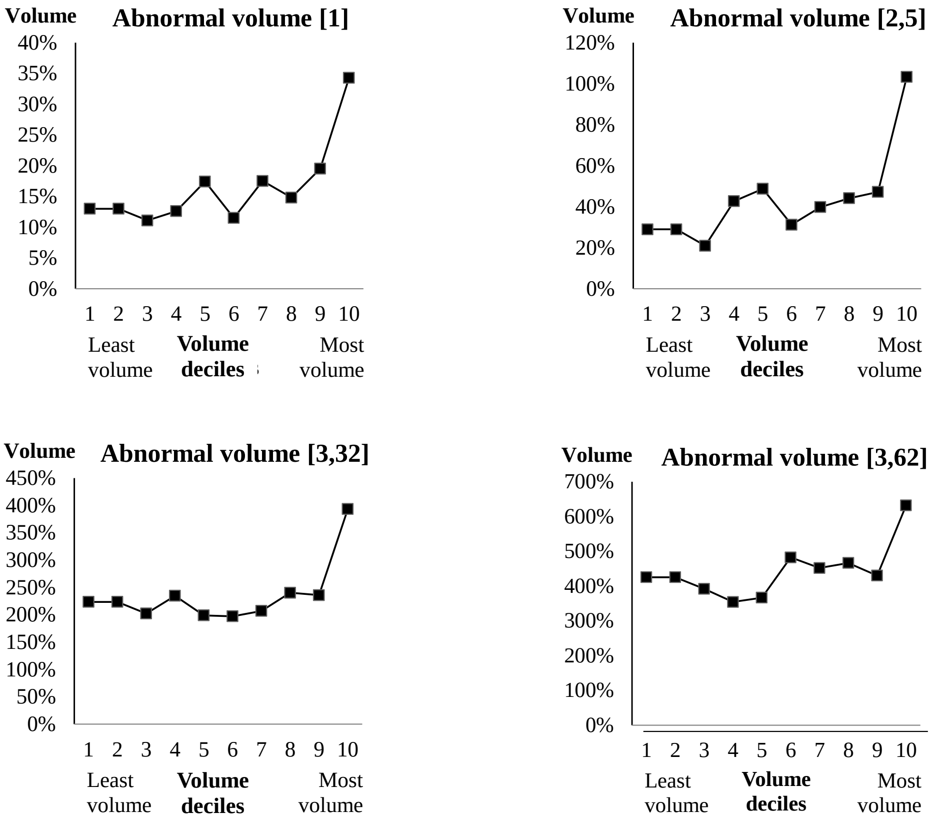


Figure 23: Firm specific news volume[177] deciles vs. decile median abnormal trading volume

In light of the univariate tests, it seems that firm specific news volume has at least a weak relationship with abnormal volume. Interestingly, we see a drastic change in the last decile[178]; a similar change was evident with firm specific news volume with abnormal returns. To some surprise, the relationship seems to be positive, and hence in contradiction with our main hypothesis. In fact, the univariate tests seem to support our alternative hypothesis suggesting that attention grabbing stocks expedite abnormal trading as is evident by the positive relationship illustrated in Figure 23. Also, it seems that the relationship is more prominent for the shorter event windows while the longer event windows seem to have slightly weaker increasing pattern. However, all the event windows show more or less stable upward movement throughout the deciles, with a radical increase in the last decile. Indeed, we suggest

[177] We use only data with news volume over 5, as otherwise too many deciles would have equal values. In other words, the first 4 deciles would have news volume value of 0, and the next two deciles would have a news volume value of 1.

[178] The reader should note that the last decile is not equal to the other deciles, and therefore a large change is natural in this case. In majority of instances, the news volume would be between 5-25 news per day. These data points are presented with deciles 1-9. Decile 10, on the other hand, captures all instances where news volume is higher than 25, in fact the highest news volume going up to 235 news items per day. (For market news volume and the LPS-score, the jump to the last decile is not drastic).

that firm specific news volume can in fact have a binomial relationship with its dependent variables. In other words, after crossing a certain threshold, a firm becomes '*attention grabber'*[179], and experiences a significant increase in its abnormal trading volume, or a decrease in its abnormal returns. However, before crossing this threshold, the dependent variables are not significantly impacted by more firm specific news. In conclusion, based on the univariate findings, we suggest that firm specific news volume has a relationship with abnormal trading volume that is positive in nature.

6.2.3 Abnormal volatility

As our final univariate test, we study the impact of our main independent variables on abnormal volatility. We will proceed in this section in a similar manner as in the previous sections: first, we plot Linearized Phrase-Structure -model sentiment against abnormal volatility; second, we plot market news volume against abnormal volatility; finally, we plot firm specific news volume against abnormal volatility. The graphs are shown below.

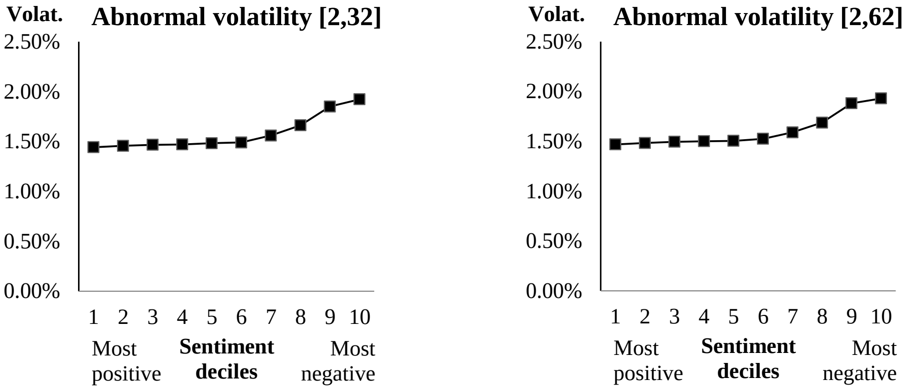


Figure 24: LPS model deciles vs. decile median abnormal volatility

Based on the univariate analysis, it seems that sentiment has the ability to forecast future abnormal volatility. As is shown in Figure 24, abnormal volatility increases monotonically throughout the deciles. However, the first deciles are relatively constant. Nevertheless, the last deciles show a clear and monotonic increase with each decile that is a strong indication of a relationship. Furthermore, the relationship is positive in nature, as we hypothesized. Indeed, our univariate findings seem to support the findings of prior literature. We infer that

[179] The firm experiences more trading due to speculators, and increased attention.

univariate tests seem to confirm our hypothesis of sentiment changes leading to abnormal volatility as noise traders react to sentiment changes with exaggerated action.

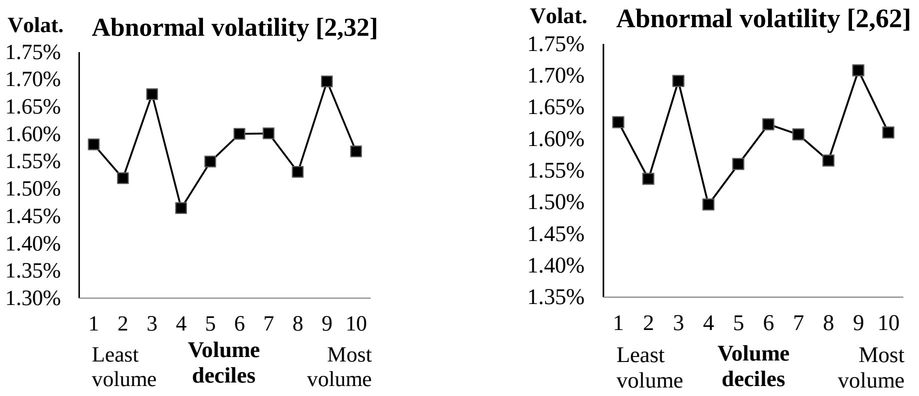


Figure 25: Market news volume deciles vs. decile median abnormal volatility

As we expected, it seems that there is no clear relationship with market news volume and firm specific idiosyncratic abnormal volatility. Indeed, we did not foresee a valid hypothesis for such a relationship, and our univariate findings seem to confirm our initial line of thinking. Nevertheless, we will study market news volume as part of our holistic media model main specification in the multivariate context. However, it seems that market news volume is not a driver of abnormal volatility.

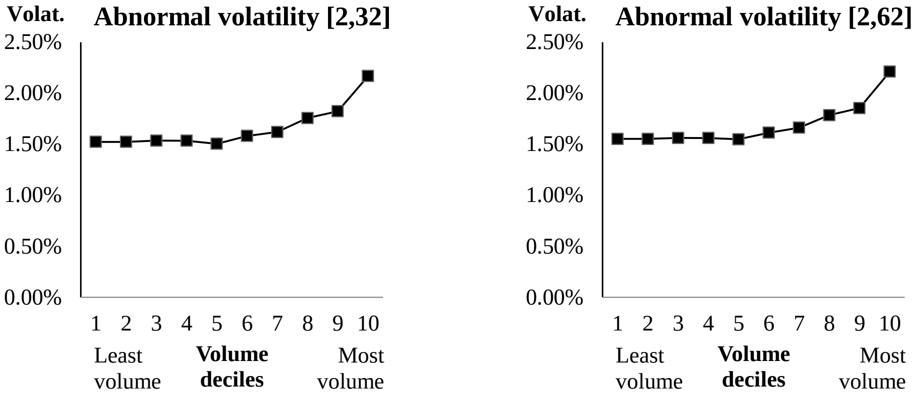


Figure 26: Firm specific news volume[180] deciles vs. decile median abnormal volatility

[180] We use only data with news volume over 5, as otherwise too many deciles would have equal values. In other words, the first 4 deciles would have news volume value of 0, and the next two deciles would have a news volume value of 1.

As with sentiment, firm specific news volume seems to have a clear monotonic relationship with abnormal volatility. Indeed, the relationship is positive in nature, and therefore in line with our hypothesis of attention grabbing stocks attracting noise trader activity. We conclude that the univariate results seem to support our hypothesis, and indicate that there is a relationship between firm specific news volume and abnormal volatility.

In general, our tests infer that the volatility impact seems to be relatively stable throughout time, and lasts for a long period of time. In other words, volatility impact persists over time, based on the univariate results. Indeed, such an empirical finding has been documented previously in prior literature (e.g., Antweiler, 2004). The next step is to analyze these relationships in multivariate context.

6.3 Multivariate tests

After studying the univariate results, we move on to multivariate specifications to better capture the relationship between our dependent variables and main independent variables while controlling for several other known factors. We will first explore the main specifications we have outlined in Section 5.4.1, and then move on to discuss the alternative specifications discussed in Section 5.4.2.

6.3.1 Main Specifications

Based on the univariate results, we are not expecting to see strong evidence of a relationship between sentiment and our dependent variables. On the other hand, we are interested to see how the relationship between news volume (market volume and firm specific volume) holds up after adding a number of control variables to the equation. To better understand the different drivers of our dependent variables, and their potential link with our independent variables, we run multivariate analysis according to the main specifications we have outlined in Section 5.3.1.

We will first explore the relationship between abnormal returns and our independent variables. From there on, we will continue by discussing abnormal volume. Finally, we conclude by looking at the impact our independent variables have on abnormal volatility.

Abnormal returns

Based on the univariate results, it seems that sentiment is unable to explain variations in abnormal returns. However, in order to better understand the relationship, and to study the effect of a holistic media model, we move to multivariate analysis, and include our two other main independent variables in the specification with a set of controls, as described in Section 5.3.1. The results of our main specification regressions are found below in Table 16.

Table 16: Multivariate specification: Abnormal return

The variable definitions are found in Section 4. Data. Discussion on methodology specification can be found in Section 5. Methodology. Specification is run using Fama-MacBeth methodology to counter time-effect present in the data. The reported coefficients, and t-statistics, are based on 21 quarters. Statistically significant (5%) t-stats are bolded. All coefficient estimates are scaled with 100 except for main independent variables that are scaled with 10,000.

	Event windows							
	[0,1]		**[1,5]**		**[2,32]**		**[2,62]**	
Sentiment	Coef.	*t-stat*	Coef.	*t-stat*	Coef.	*t-stat*	Coef.	*t-stat*
LPS	-0,03	*-0,46*	-0,27	***-2,02***	-0,42	*-0,86*	-1,21	*-1,27*
Wordcounts								
Finance dictionary	0,61	*0,72*	-0,21	*-0,15*	-5,54	*-0,75*	-14,86	*-0,99*
H4N dictionary	0,14	*0,20*	-1,13	*-0,76*	-0,78	*-0,13*	-10,07	*-0,80*
Main Independent Variables*								
Market news volume	-0,01	*-0,47*	0,20	***2,29***	0,89	***3,11***	1,68	***2,87***
Firm news volume	0,00	*-0,02*	-0,46	*-1,02*	-2,45	*-1,09*	-2,89	*-0,72*
Control Variables*								
Size	0,12	***3,31***	-0,11	*-1,20*	-0,86	*-1,49*	-1,73	*-1,73*
Book-to-market	-0,85	***-2,66***	0,75	*1,19*	2,28	*1,05*	3,51	*1,16*
Momentum								
[-4,-1]	-2,27	***-4,32***	-6,25	***-5,55***	-13,70	***-5,87***	-15,06	***-3,91***
[-34,-4]	-0,86	***-6,31***	-2,34	***-4,69***	-7,94	***-2,94***	-6,35	*-1,50*
[-255,-34]	-0,10	*-0,90*	-0,04	*-0,13*	0,19	*0,15*	1,19	*0,51*
Share turnover	0,14	*1,61*	0,71	***2,00***	4,50	*1,98*	8,59	***2,19***
SUE	-0,38	*-0,33*	-4,87	*-0,96*	18,05	*0,68*	37,79	*0,73*
Abnormal volatility	5,75	*1,81*	-14,23	*-1,30*	-83,22	*-1,59*	-109,57	*-1,05*
Institutional ownership	0,14	*1,10*	-0,34	*-0,91*	-2,26	*-1,09*	-4,28	*-1,07*

* Coefficients and t-stats of Main Independent and Control variables refer to LPS regression results. Word count regressions have been run separately from the LPS regression.

Sentiment

As we have shown in the previous sub-section: Section 6.1., our sentiment estimation methodology is superior to the prevalent dictionary based word counts. However, even though Linearized Phrase-Structure -model fares better, the results are not promising. Based on the multivariate analysis, in combination with the univariate analysis, we suggest that in light of our empirical findings, different sentiment estimates cannot be used to forecast abnormal returns. In fact, even with the most promising event window: the event window of [1,5], the coefficient signs are not consistent within the 21 quarters. Hence, the reliability of the estimate is questionable, and any inference drawn from it should be judged with a grain of salt.

In spite of not being able to draw exact inferences from the results, we do suggest that the qualitative nature of our results offers some light to the role of qualitative texts in financial markets. It seems that the hypothesis suggesting that qualitative texts hold informational content has some support from our results.[181] As our results show, the different sentiment estimate coefficient signs are all negative throughout the event windows.[182] Therefore, they show a pattern of underreaction to the increase in negativity concerning a company: the negative news are disseminated slowly into the stock price resulting in a negative abnormal return drift. Moreover, the coefficient signs grow monotonically with event window horizons, illustrating that the dissemination of information escalates as time passes on and agents are able to analyze all the information.[183] The suggested qualitative link is in line with previous studies on content analysis in the field of finance (for more information on prior findings, we refer the reader to Section 2.3).

All in all, we conclude that the magnitude of negativity in sentiment seems to have limited impact on abnormal returns. We cannot rule out the hypothesis suggesting that tone has an impact on investors through framing; however, it seems that on the aggregate level, the markets are not substantially affected by the framing bias. Also, we cannot take a definite stand on the topic of qualitative text's informational content, but our results do suggest that the impact of informational content would dominate the reaction of investors over tone, when considering the impact of qualitative texts on abnormal returns as is evident by the underreaction pattern. However, neither of the effects can be reliably used to forecast abnormal returns. In conclusion, our results suggest that the level of sentiment negativity is not a factor of abnormal returns: either markets are efficient in disseminating information in qualitative texts, or there is no information in qualitative texts. Furthermore, our results

[181] The competing hypothesis would be that news have an impact on readers through tone - the theory behind this hypothesis is based on the effects of framing. More discussion on Section 3 and Section 2.

[182] The exceptions are the 1-day results for the two dictionaries. However, as we simply suggest a very weak qualitative inference based on our results, we do not consider this discrepancy as important.

[183] An alternative explanation is that 'bad model' problem drives the increase in coefficients, as abnormal return estimation is more prone to the aforementioned problem with longer event horizons. However, the pattern exists with alternative abnormal return calculation methods in alternative specifications.

suggest that tone has no significant predictability over abnormal returns on aggregate market level.

In light of our findings, we argue that the findings of prior literature are characterized by spurious regularities as suggested by many of the EMH proponents in the context of market anomalies (e.g., Fama, 1991, 1998; Schwert, 2003; Malkiel, 2003).[184] Our study is, to our knowledge, the most comprehensive study in finance in the field of content analysis. Whereas other studies have focused on specific qualitative text sources (i.e., 10-ks, earnings announcements, specific news journals, etc.,), our study has analyzed a wide cross-section of qualitative texts. Furthermore, whereas prior studies have focused on one or two arbitrary event windows, we have utilized several event windows to examine the impact of sentiment on abnormal returns. Moreover, as we will later on describe, we have utilized several different methodologies in estimating abnormal returns, and sentiment, to avoid *'bad model'* problems. In conclusion, our findings provide robust empirical evidence that suggests that there is no significant link between abnormal returns and sentiment. We conclude by suggesting that the findings of extant literature are mainly spurious regularities that are a result of data dredging with specific sources of qualitative texts, and the use of particular event windows and methodologies, to maximize results.

Market news volume

Our results for market news volume hold much more promise than the sentiment estimates. Indeed, in line with prior literature (e.g., Hirsleifer et al., 2009), market news volume shows opposite signs for short-term event windows vis-à-vis long-term event windows as hypothesized by limited attention theory. One day after the event, the relation with the variable to abnormal returns is negative, while the longer event-windows show a positive relationship: an indication that the variable proxies for limited attention and underreaction to information. Moreover, the following day relationship with abnormal returns is not statistically significant, as hypothesized by limited attention theory: as information is not compounded to the stock price immediately due to distraction caused by other news, the following day relationship with abnormal return should not be statistically significant. As information begins to disseminate later on, the relationship with abnormal returns turns to significant for the longer event windows.

[184] See Section 2.1., for more discussion on efficient market hypothesis, and Section 2.2., for discussion on Behavioral finance.

Whereas Hirsleifer et al., (2009) used only the number of earnings announcements as a proxy of distraction, our sample has counted also news texts as proxy for market wide distraction, and the results remain the same. Moreover, our study provides support for the hypothesis using several different event windows. We conclude that our results provide strong backing for Hirsleifer et al., (2009) hypothesis that simultaneous market news distract investors. This forces investors to divide their attention, resulting in a slow incorporation of information into stock prices: a drift.

Our results are robust to several different methodologies discussed later on in alternative specifications in section 6.3.2. Furthermore, whereas with sentiment estimate coefficients, the signs of the coefficients were changing from quarter to quarter, market news volume coefficient signs remain monotonic throughout quarters. With statistical significance, and monotonic qualitative results for coefficient estimates throughout time clusters, our results appear robust.

Firm specific news volume

The results relating to our last independent variable: firm specific news volume, fall into the same category with sentiment estimates. The results are not statistically significant, but provide some insight into the qualitative nature of the relationship between firm specific news volume and abnormal returns. As we hypothesized, firm specific news volume has a negative relationship with abnormal returns. The aforementioned finding is in line with prior literature, and limited attention hypothesis. Indeed, as limited attention theory would suggest: the more news a firm has, the more attention it will attract, and therefore the more efficiently the information will be incorporated into its stock price. In other words, firm specific news volume exhibits the opposite relationship compared to market news volume with abnormal returns. The empirical findings support our hypothesis qualitatively.

Majority of our control results are in line with our expectations. As we are using matching portfolios based on size and book-to-market as benchmark returns, we would not expect size and book-to-market to have significant relationship with abnormal returns.[185] Momentum variables exhibit a reversal relationship with abnormal returns. The finding is not in line with the original anomalies discussed in Section 2.1., that state that on short-term previous winners should outperform previous losers, and on long-term stocks exhibit reversal. However, recent

[185] The exception of event window [0,1] remains a puzzle; potentially indicating that the variables capture the effect of an omitted variable that has an impact on the dependent variable in the event window in question.

empirical findings (e.g., Engelberg, 2008; Tetlock et al., 2008; Loughran and McDonald, 2011) support our findings which seem to illustrate a reversal effect in the short-term[186]. Our other controls seem to lack significance but are qualitatively mainly in line with our expectations.

Abnormal volume

Based on the univariate results, it seems that different sentiment estimates might explain abnormal volume variations. In order to pinpoint the exact relationships, we move to multivariate analysis, and include all of our main independent variables in the specification with a set of controls, as described in Section 5.3.1. The results of our main specification regressions are found below in Table 17.

Table 17: Multivariate specification: Abnormal volume

The variable definitions are found in Section 4. Data. Discussion on methodology specification can be found in Section 5. Methodology. Specification is run using Fama-MacBeth methodology to counter time-effect present in the data. The reported coefficients, and t-statistics, are based on 21 quarters. Statistically significant (5%) t-stats are bolded. Main independent variable coefficient estimates are scaled with 100.

	Event windows							
	[1]		[2,5]		[3,32]		[3,62]	
Sentiment	Coef.	*t-stat*	Coef.	*t-stat*	Coef.	*t-stat*	Coef.	*t-stat*
LPS	0,08	*1,39*	0,12	*0,62*	1,07	*0,98*	0,89	*0,57*
Wordcounts								
Finance dictionary	1,58	***2,51***	1,44	*0,49*	3,59	*0,27*	16,15	*0,76*
H4N dictionary	1,04	*1,99*	-1,11	*-0,52*	-4,89	*-0,48*	5,59	*0,34*
Main Independent Variables*								
Market news volume	0,06	***2,14***	0,31	***2,76***	1,31	***2,26***	2,04	***2,44***
Firm news volume	0,87	***5,73***	1,50	***4,83***	3,77	*1,75*	4,89	*1,14*
Control Variables*								
Size	-0,14	***-5,04***	-0,36	***-3,54***	-1,57	*-1,66*	-1,68	*-0,88*
Book-to-market	0,28	*1,54*	1,36	***2,03***	7,70	***2,19***	17,95	*2,78*
Momentum								
[-4,-1]	-1,20	***-3,77***	-2,44	***-2,74***	-18,11	***-5,43***	-26,40	*-4,78*
[-34,-4]	-0,60	***-3,92***	-2,57	***-4,32***	-15,98	***-4,29***	-16,16	*-4,50*
[-255,-34]	0,03	*0,40*	0,18	*0,66*	2,72	*1,78*	4,32	*2,11*
Share turnover	-0,22	***-2,70***	-0,53	*-1,86*	-3,94	***-2,42***	-8,01	*-2,40*
SUE	3,08	*1,04*	9,43	*0,83*	63,37	*1,47*	26,48	*0,48*
Abnormal volatility	-8,76	***-2,10***	-45,84	***-2,87***	-244,46	***-2,63***	-366,98	*-2,26*
Institutional ownership	0,14	*0,83*	0,44	*0,74*	2,61	*0,75*	7,38	*1,25*
Abnormal market volume	0,50	***19,87***	0,39	***10,70***	0,19	***4,30***	0,10	***4,13***

* Coefficients and t-stats of Main Independent and Control variables refer to LPS regression results. Word count regressions have been run separately from the LPS regression.

Sentiment

[186] In addition, we recognize that it is possible that past stock returns lead to changes in sentiment. Therefore, correlation between these two variables is likely, and the coefficients of both may be somewhat biased.

Even though univariate tests held some promise for the sentiment variables, sentiment is unable to forecast abnormal trading volume changes based on the multivariate results. With the exception of Loughran and McDonald (2011) dictionary result for event window [1], all the variables are statistically insignificant. Moreover, even with the result of event window [1], the coefficient signs are not monotonic throughout the 21 quarters; hence, casting doubt on the potential inferences drawn from the event window [1] result. Thus, it appears possible that sentiment simply mimics some of the control variables, therefore showing promising results in univariate tests.

Nevertheless, our results do offer support qualitatively to the hypothesis that information content dominates investor reactions more than tone. As with returns, limited attention and underreaction hypothesis are supported by the empirical findings in abnormal volume specification. As we can see, coefficients are all positive throughout the event windows, and increase in magnitude when moving to longer event windows - as is expected by the limited attention hypothesis that suggests underreaction to new information.[187] As a result, our empirical findings offer weak support for the branch of prior literature that has found underreaction to qualitative text information; in other words, our findings seem to indicate an increase in abnormal volume post-event (e.g., Loughran and McDonald, 2011).[188]

Market news volume

In light of the univariate tests, we had lower expectations for market news volume than for sentiment variables. However, in the multivariate specification, market news volume seems to explain abnormal volume variations. In fact, our results suggest that our initial hypothesis for the relationship between abnormal volume and market news volume was correct: market news volume proxies for distraction and causes markets to underreact to new information; hence, resulting in abnormal volume throughout all the event windows[189]. Moreover, interestingly, we can see that the coefficient magnitude increases with longer event periods. We infer that as

[187] The exception here is the Harvard negative dictionary. However, as we have established in previous sections, Loughran and McDonald demonstrated that Harvard Dictionary misclassifies words in financial context approximately 75% of time. Moreover, our own sentiment estimate benchmarking provided proof that H4N dictionary is inferior to the two other methods. Therefore, we assign the discrepancy as an outcome of noise in the H4N variable.

[188] For more information on prior literature findings with financial metrics, see Section 2.3.

[189] An interesting small deviation from our initial hypothesis is the positive significant relationship in the succeeding day. We did not exclude the possibility of a significant relationship, but the sign of the coefficient is a surprise. However, the magnitude of the coefficient is so low that we do not consider this finding as contradictory to our theory building based on our findings.

information disseminates with time, agents are able to act on that information, causing prolonged elevated levels of trading.

Our empirical results offer strong support for Hirshleifer et al. (2009) distraction hypothesis, and the underlying limited attention theory. Furthermore, our findings with abnormal volume are in line with our findings with abnormal returns. In other words, our empirical findings seem to suggest that there is informational content in qualitative texts, and that text is incorporated into financial metrics with delay due to underreaction.[190]

Firm specific news volume

Based on the univariate tests, firm specific news volume seemed to have a relationship with abnormal volume variations. As expected, firm specific news volume seems to explain abnormal volume changes also in the multivariate context. However, the statistical significance drops dramatically when moving to the longer term event windows. Therefore, it seems that firm specific news volume has explanatory power over abnormal volume, but only in the short-term – a finding supported by the univariate tests, and in line with our hypotheses.

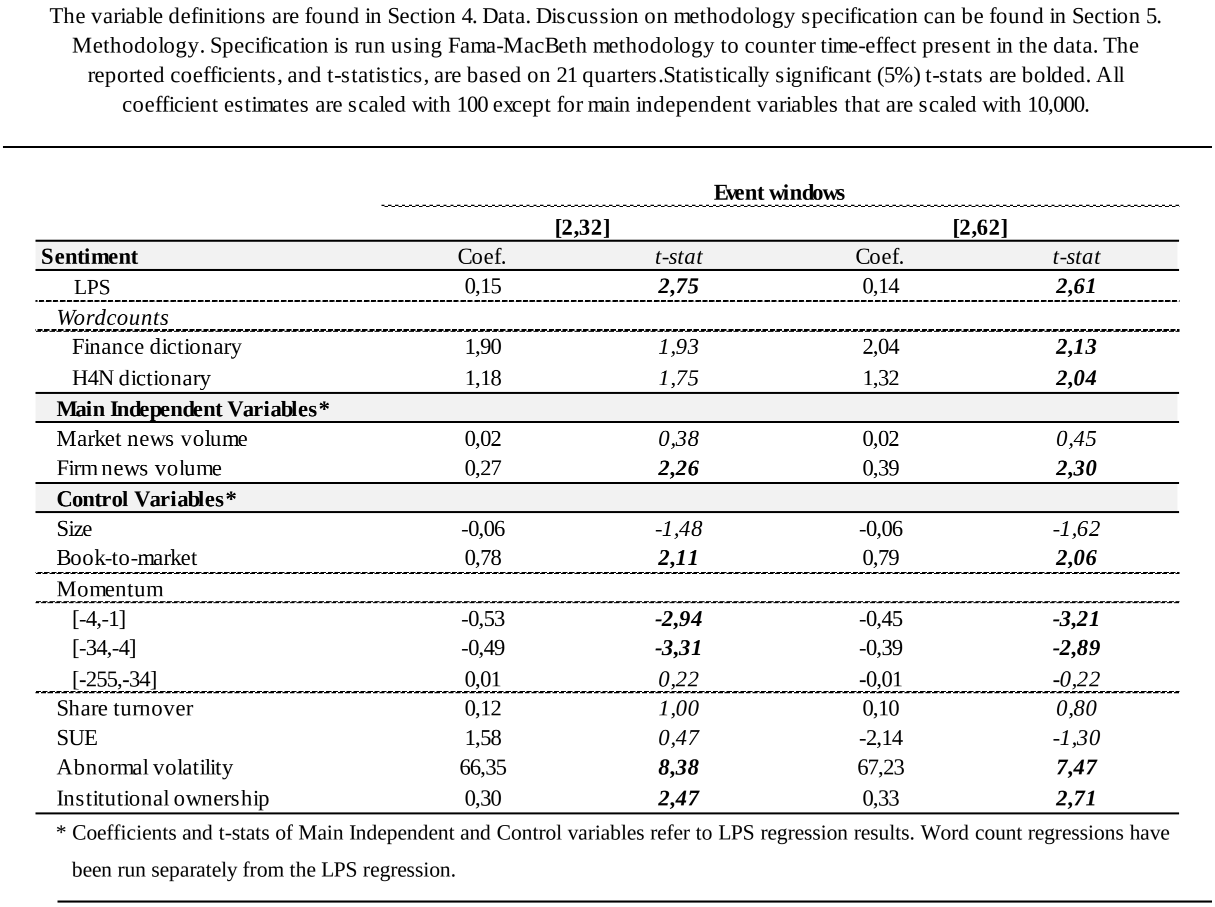
To some extent, the empirical findings relating to firm specific news volume and abnormal volume are surprising. Contrary to our primary hypothesis, firm specific news volume magnitude increases as event window lengths increase. These results lack statistical significance for the longer event windows, so the results could be interpreted as directly supporting our hypotheses. Nevertheless, our findings seem to support the hypothesis that attention grabbing stocks get traded more actively throughout the event windows. In contrast to our primary hypothesis that stated that the relationship should weaken with longer event windows as more attention should mean quicker dissemination of information and therefore less abnormal trading activity in the long run, the abnormal volume persists and even increases with longer event windows. However, as the longer event windows are not statistically significant, we suggest that attention grabbing stock experience abnormal trading activity close to the event day, but that effect disappears as time goes on. In other words, attention grabbing stocks abnormal volume levels do not experience a prolonged elevated level ex-post the event.

[190] However, as our results are statistically weak, we can only provide weak qualitative results, and our inferences should be judged in that context.

Abnormal volatility

In light of the univariate tests, it seems that different sentiment estimates hold great promise in explaining future idiosyncratic abnormal volatility. In order to specify the relationship, we move to multivariate analysis, and include all of our main independent variables in the specification with a set of controls, as described in Section 5.3.1. The results of our main specification regressions are found below in Table 18.

Table 18: Multivariate specification: Abnormal volatility



The variable definitions are found in Section 4. Data. Discussion on methodology specification can be found in Section 5. Methodology. Specification is run using Fama-MacBeth methodology to counter time-effect present in the data. The reported coefficients, and t-statistics, are based on 21 quarters.Statistically significant (5%) t-stats are bolded. All coefficient estimates are scaled with 100 except for main independent variables that are scaled with 10,000.

	Event windows			
	[2,32]		[2,62]	
Sentiment	Coef.	*t-stat*	Coef.	*t-stat*
LPS	0,15	***2,75***	0,14	***2,61***
Wordcounts				
Finance dictionary	1,90	*1,93*	2,04	***2,13***
H4N dictionary	1,18	*1,75*	1,32	***2,04***
Main Independent Variables*				
Market news volume	0,02	*0,38*	0,02	*0,45*
Firm news volume	0,27	***2,26***	0,39	***2,30***
Control Variables*				
Size	-0,06	*-1,48*	-0,06	*-1,62*
Book-to-market	0,78	***2,11***	0,79	***2,06***
Momentum				
[-4,-1]	-0,53	***-2,94***	-0,45	***-3,21***
[-34,-4]	-0,49	***-3,31***	-0,39	***-2,89***
[-255,-34]	0,01	*0,22*	-0,01	*-0,22*
Share turnover	0,12	*1,00*	0,10	*0,80*
SUE	1,58	*0,47*	-2,14	*-1,30*
Abnormal volatility	66,35	***8,38***	67,23	***7,47***
Institutional ownership	0,30	***2,47***	0,33	***2,71***

* Coefficients and t-stats of Main Independent and Control variables refer to LPS regression results. Word count regressions have been run separately from the LPS regression.

Sentiment

As we expected, sentiment estimate changes have a positive relationship with abnormal volatility variations. As with our benchmarking test, Linearized Phrase-Structure -model outperforms the two other sentiment estimation methodologies when measured with statistical significance. Similarly, Loughran and McDonald's (2011) finance dictionary outperforms H4N dictionary. Also, an interesting finding is that dictionary based word count coefficients seem to overestimate the impact of sentiment on volatility vis-à-vis Linearized Phrase-

Structure -model. We suggest that the underlying cause is the noise present in the sentiment estimates derived in the less precise word count methodology.

When looking at the two different event windows, it is noteworthy to state that the length of the event window seems to have little to no impact with the magnitude of the change in the sentiment estimate coefficient. Also, the pattern is, to some extent, visible with the other independent variables. Therefore, we infer that the volatility effect persists over long periods of time without perishing, or increasing. In conclusion, our empirical findings are in line with that of prior literatures[191], linking sentiment changes with future abnormal volatility.

Market news volume

As we did not have a clear hypothesis for market news volume, we did not expect to find a significant relationship between the variable and abnormal volatility. Indeed, univariate tests gave the first indication that market news volume does not have a relationship with abnormal volatility. The multivariate tests confirm this interpretation. Besides being statistically insignificant, the coefficient of market news volume is marginal. Indeed, it seems that market news volume has no relationship with a given firm's abnormal volatility.

Firm specific news volume

Contrary to market news volume, firm specific news volume seems to have a relationship with abnormal volatility based on the univariate tests. Indeed, multivariate tests confirm this observation. As we hypothesized, firm specific news volume has a positive relationship with abnormal volatility. We interpret this finding to confirm our hypothesis that noise traders are attracted to trade on a stock that has high visibility in the media. Therefore, noise traders elevate the volatility levels of a given firm's stock with their trading behavior. The hypothesis is supported to some extent by the findings in abnormal trading volume.

All in all, our empirical findings are in line with prior literatures findings. We infer that volatility is related to sentiment changes, and to firm specific news volume. Our results hold for both of two event windows we have employed, and for several different return methodologies. Therefore, we conclude that our findings provide strong evidence in support of the link between the aforementioned variables and abnormal volatility.

[191] For more discussion, see Section 2.3.

6.3.2 Alternative specifications

As we discussed in Section 5.4.2, we run several alternative specifications to pinpoint the exact relationships underlying our main research variables, and to be as robust as possible with our results. Also, we conduct an additional study employing a new dependent variable: aggregate market index returns. We will briefly describe the results of the aforementioned in this sub-section.

First, we will describe the results of our main specification with alternative abnormal return methodologies. Second, we will portray the impact that alternative main variable definitions have on our results. Third, we will discuss the impact that additional control variables have on our results. Finally, we will study the impact of our main specification for abnormal returns, but with aggregate market index return acting as the dependent variable.

Alternative abnormal return definitions

To counter any critique towards *'bad model'* problems, and to make our study robust in terms of methodology, we run several different abnormal return specifications as discussed in Section 5.4.2.

First, we use raw returns instead of abnormal returns in line with Tetlock (2007). We find an increase in coefficient estimates and statistical significance in all our main variables: sentiment, market news volume and firm specific news volume. Therefore, we conclude that linking raw returns to our variables is less demanding than to establish the link between abnormal returns. However, the economic significance of such findings has limited value.

Second, we use two different abnormal return benchmarks: value weighted index returns and Fama and French three-factor model returns. For value weighted index returns, our results improve marginally. However, for Fama and French three-factor model returns our results decrease substantially. However, market news volume's impact remains similar to our main specification, adding needed robustness to the finding.

Third, we switch from using BHARs into using CARs when calculating our abnormal returns. As we expect, we see drops in our coefficient estimate magnitudes. However, statistical significances are not impacted by the methodology switch. Therefore, the inferences we have drawn from our findings remain the same.

We conclude by stating that our results are fairly robust to methodological choices in terms of abnormal returns. There are methodologies that would yield better results vis-à-vis our main specification methodology. However, there are also methodologies that would result in worse results as the ones we are now reporting with our main specification.

Alternative main variable definitions

As discussed in Section 5.4.2, we run alternative definitions of our main study variables. To see whether or not seasonal declining trend in news volume impacts our results, we run our main specifications with a standardized version of market and news volume variables. Also, we test the abnormal aspect of both variables by deducting average past volume. We find mixed impacts on our coefficients and their statistical significance. However, there are no apparent systematic pattern changes. Therefore, in line with prior literature, we conclude that seasonal trend is not driving our results, and standardization is not necessary.

Secondly, we test our alternative metrics of sentiment, as specified in section 5.2.4. The results show some variance from sentiment metric to sentiment metric. However, as the differences are not major, after several experiments we choose to stick to our original LPS method of aggregating sentiment of all media for one day.

Additional control variables

To mitigate the possibility of an omitted variable impacting our results, we run several alternative specifications with additional controls.[192] The variable descriptions are listed in Section 4.2.2, and the specifications in Section 5.4.2. The findings of our alternative control specifications are listed below.

Impact of analysts

To test the impact analysts have on information dissemination and on the subsequent impact on our dependent variables, we include the number of analysts following a company as well as analyst dispersion to our main specifications. After the inclusion, our results remain similar to main specifications. However, analyst variables subsume the significance of share turnover, while also turning the coefficient into negative. We conclude that the variables seem to proxy for analyst disagreement that conveys belief dispersion in the market previously captured by

[192] Additional controls are added to the main specifications described in Section 5.3.1. We do not run the alternative specifications previously introduced with additional controls.

share turnover, while turnover moves to capture liquidity effects. In such case, the signs appear rationale. However, as noted, both turnover and analyst variables remain insignificant.

Calendar effects

As we are using matching portfolios as our benchmark returns in our main specifications, we are implicitly controlling for calendar effects.[193] Also, our abnormal volume specification takes into account abnormal market trading volume that implicitly controls for calendar effects in the same manner as benchmark returns. Nevertheless, as described in Section 4.2.2., we include calendar variables into our main specifications. We include January, Monday, and end-of-month dummies into our specifications. The result is not surprising: all of the dummies are insignificant, and the results are practically unchanged.

Dividend effects

In line with Li (2006), we control for the potential impact of dividends on our abnormal return main specification. We do not perform this check for abnormal volume or abnormal volatility, as there is no prevalent hypothesis that dividends should bear an impact on these variables. After the inclusion of dividends, we do not see significant changes in our other variables. Dividends seem to have qualitatively a positive relationship with abnormal returns, but lack statistical significance. We conclude that dividends are not impacting our results.

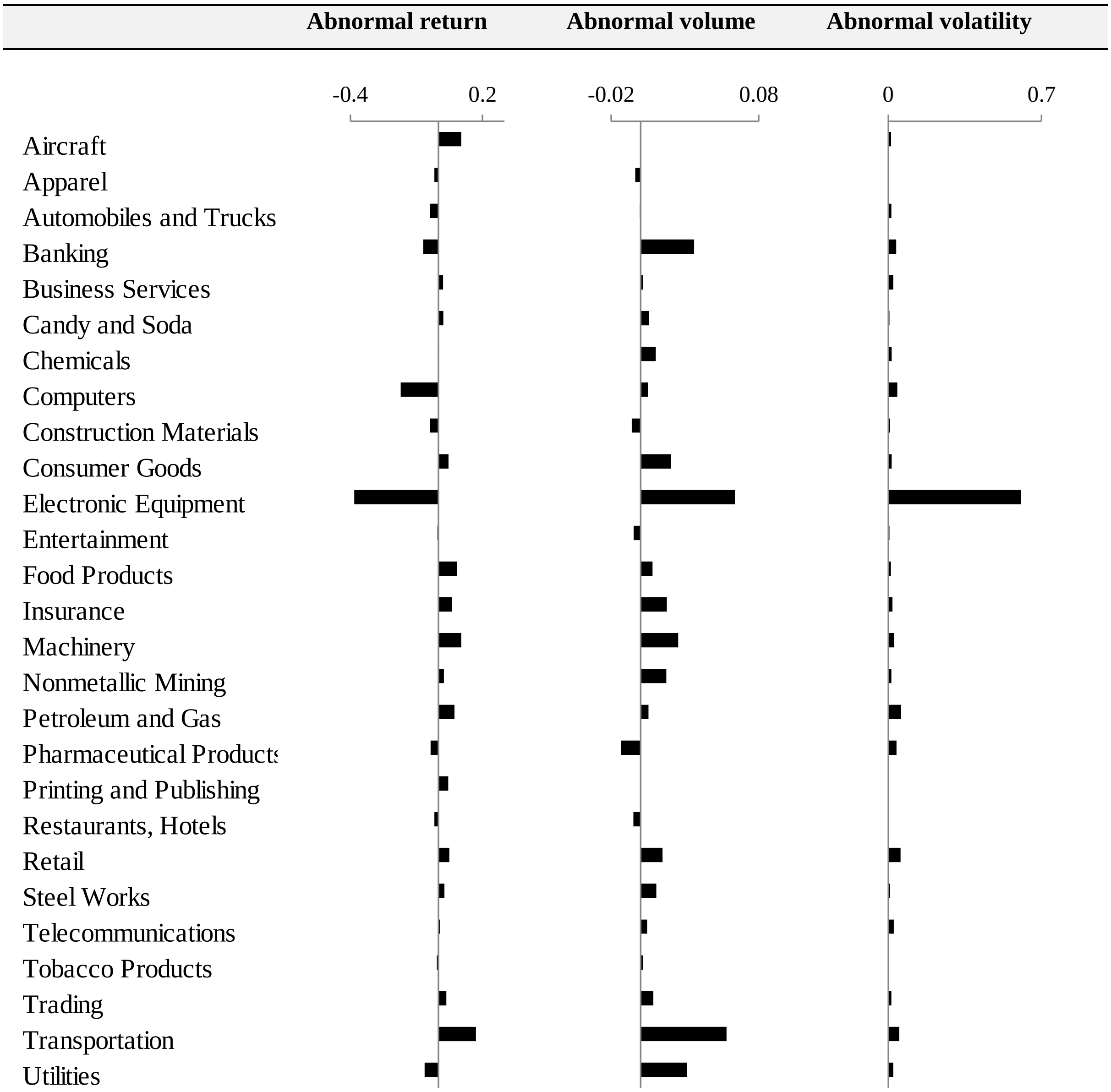
Industry specific effects

To study the possible impact of industry effects, we start with a simple analysis by calculating the average values of all of our dependent variables for all industries. We calculate these averages in order to see whether any industry would have significantly different values compared to others and thus be impacting our results. For example, we hypothesize that during the financial crisis the '*banking'* and *'trading*' industries could have significantly differing values.

[193] The matched portfolio that acts as a benchmark return is also impacted by calendar effects.

Table 19: Stock performance metrics by industry[194]

Average stock performance metrics over the whole time period for each industry. Only average values for 60-day event windows shown (average values are similar regardless of event window). Return and volatility average figures have been multiplied by 1 000.



The data clearly shows that most industries have close values to each other. Contrary to our presumption, *banking* and *trading* are also well in line with the average values. The only larger outlier appears to be *electronic equipment* industry that consists of three companies: Cisco, Intel and Qualcomm. We do recognize that this industry may be leading some part of our results. However, as the industry is relatively small: only 3/100 companies, we do not

[194] Only values for 60-day event windows shown (average values similar regardless of event window

consider that this would be a significant issue. Also, as the industry is small in terms of the number of companies, it is possible that only one company performing extraordinarily could be driving the finding.

As described in Section 4.2.2, we next incorporate industry dummies to take into account the potential driving impact of an industry related factor. After categorizing our firms based on Fama and French (1997) classifications, our sample is divided into 28 different Fama and French industry categories. Hence, we include 27 industry dummies with one category acting as the omitted category.

We run regressions for the whole data sample both with and without industry dummies. Most industries do receive significant coefficients in our regression. The main coefficients for our main specification with and without industry dummies are reported in the Table 20.

Table 20: Alternative specification: Included industries

The variable definitions are found in Section 4, Data. Discussion on methodology specification can be found in Section 5, Methodology. We run a regression with SVM-sentiment for the full time period at once, with and without industry dummies. The table reports the LPS-sentiment coefficients. Coefficient values for returns and volatility have been multiplied by 1 000. Results differ slightly from previous as we limited the number of control variables only to those that have shown significant coefficients in previous regressions.

	Industries included		**Industries excluded**	
	Coefficient	*t-stat*	Coefficient	*t-stat*
Return [0,1]	-0.397	-0.769	-0.379	-0.723
Return [1,5]	-1.952	-1.949	-2.027	-1.995
Return [2,32]	-0.735	-0.301	-1.609	-0.649
Return [0,4]	-5.417	-1.608	-6.902	-2.021
Volume [1]	110.053	3.103	115.404	3.210
Volume [2,5]	533.992	5.188	538.542	5.188
Volume [3,32]	2.317	5.752	2.659	6.526
Volume [3,62]	2.450	4.497	2.972	5.373
Volatility [2,32]	5.843	24.024	5.881	23.627
Volatility [2,62]	5.926	24.870	5.964	24.408

In these results we do not notice significant variation in the coefficients or t-statistics for our main variables (t-stats and coefficient remain typically very similar). Thus this analysis leads us to believe that industries are not driving our results.

Finally, we test the coefficients of some of the largest companies with the largest news volumes: Apple, Google, Ford, AT&T and Citi. As the sample size would be too small for a quarterly Fama-McBeth regression, we only split the sample to three parts[195]. Similarly to our main regression, we notice that the coefficients for these companies are in line with main regression results. This also indicates that the largest companies are not driving the results, or that adding smaller companies with less news does not distort the results.

Based on these tests, we conclude that industries or individual companies do not appear to be driving our results. However, we recognize that adding the industry dummies can add some explanatory power to our regressions.

Additional dependent variable study: market index returns

After studying the impact on firm-specific returns, we wish to also test for the possibility that a market sentiment could have an impact on index returns. For example, firm-specific news could be well studied by analysts and therefore there might not be a visible impact, but on the overall market level the news could still carry some value. There are two reasons we hypothesize that impact of the sentiment could show in the performance of the market index more significantly than for individual firms. First, the aggregated score from hundreds of daily news varies significantly less from day to day than the firm-specific sentiment. Second, there could be market-wide effects that may not be visible if reading company-level texts (e.g. news about the economy that have only marginal impact on a specific firm). However, when all market news are aggregated and the impact on the market index is studied, these effects could become visible.

To test the hypothesis that the market sentiment, and the development of the market index, could have a connection, we pool all news in the market, and contrast this to the market index. Due to the short time period, we divide our sample only into three parts[196] and test various specifications of the SP100 sentiment and stock performance metrics. We include in each regression Market news volume, and control also for momentum, past abnormal volatility, and average market share turnover and average market SUE. We test similar combinations of returns and predicting factors as in the firm-specific context.

[195] We split the sample into three equally sized parts, the first part containing data from Jan-06 to Oct-07, the second from Oct-07 to Jun-09, and the third from Jun-09 to Mar-11. This lets us have a sufficiently large sample size of >440 for each time period. Running regressions for the index by quarter would not be sensible due to too small sample size.

[196] Similarly as for individual companies (see previous footnote), we split the sample to three equally sized parts.

Table 21: Index stock performance regressions- coefficient significance

Dependent variables are shown with their time period at the end (e.g. “Return4 = Return for period [1,5]. For sentiment, we also test whether a longer-term average would lead to more significant results: past 4-day average sentiment = “LPS_4d”, past 30-day average sentiment = “LPS_30d”.

		Independent variable		
		LPS	LPS_4d	LPS_30d
Dependent variable	Return1			
	Return4			
	Return30	*	*	**
	Return60	**	**	*
	Volume1			
	Volume4			
	Volume30			
	Volume60			
	Volatility30			
	Volatility60	*	**	**

* Significant after regression

** Significant after splitting sample into three pieces

For most of the combinations that we try, our coefficients vary between the three time periods. However, we cannot rule out the possible significance of the following regression results.

Table 22: Index stock performance regression - coefficients

Dependent	Main independent	Coefficient	T-stat
Return [2,62]	LPS	-0.256	-2.65
Return [2,62]	4-day LPS	-0.471	-2.749
Return [2,32]	30-day LPS	-0.936	-2.621
Volatility [2,62]	4-day LPS	0.02	2.128
Volatility [2,62]	30-day LPS	0.063	3.052

For all of the regressions above, our coefficients keep the same sign during the 3 time periods and the t-statistics are significant. As pointed out, we try a number of combinations to arrive to these results, and the question of data mining is a relevant concern in this case. However, we do recognize that it could be possible that sentiment would have some predictive power over market index returns or volatility. To properly test this, a longer time frame of data would naturally need to be applied, and the exercise should preferably be combined by adding multiple market indexes to test if this possibly a peculiarity of the S&P index. We also illustrate our results below in Figure 27: Sentiment vs. stock performance.

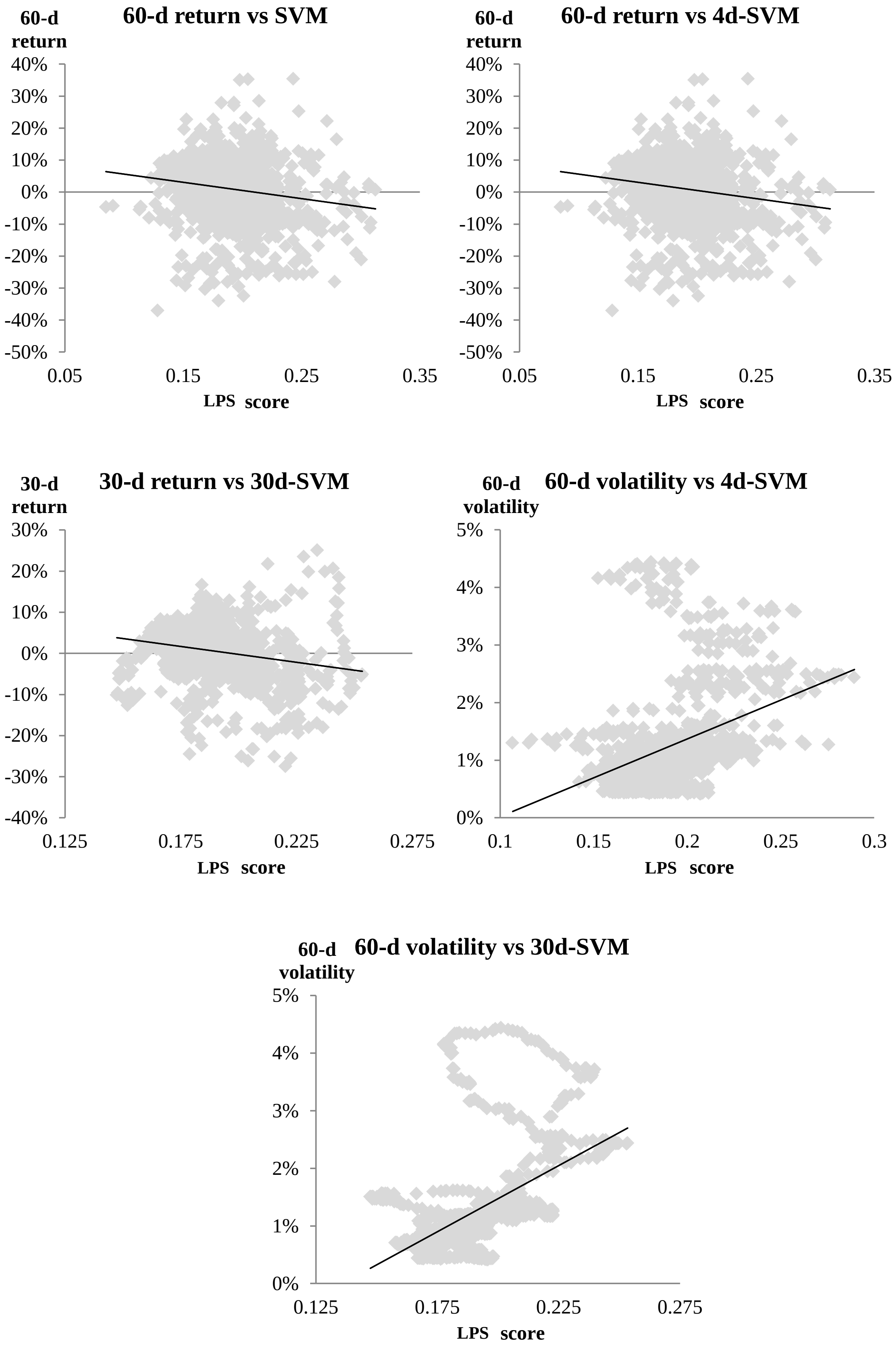


Figure 27: Sentiment vs. stock performance

6.4 Discussion on the implications of results

Our results have several interesting implications on financial theories and for future research. We start by discussing the impact on sentiment methodologies. Next, we explain how behavioral finance theories, especially limited attention, fit in line with our results. Finally, we discuss the implications on the possible link between media metrics and stock performance.

6.4.1 Impact on sentiment methodologies

Prior literature has debated on the complexity of methodology used in estimating investor sentiment, as we have discussed in Section 2.3. Scholars have proposed compelling arguments on the behalf of simple and rudimentary methodologies for estimating investor sentiment, reasoning mainly that transparency, objectivity and replication of results would suffer with the use of more advanced and complex methods (e.g., Tetlock et al., 2008). However, some researchers have attempted to employ more sophisticated methods; alas, with unexciting results (e.g., Antweiler and Frank, 2004, 2006; Das and Chen, 2006). The failure of advanced methods has been allocated to the scope of the qualitative text sources, and the limited sample sizes used in the studies in question. However, another explanation is that the findings documented with more naïve methods are in fact spurious. Our study utilizes a wide cross-section of qualitative texts with a fairly large sample, and can therefore offer improved interpretations over previous studies with issues relating to sample size and scope.

In section 6.1.1, we benchmark the existing prevalent methodology of vector word counts against our proposed methodology: the Linearized Phrase-Structure -model. The results are clear: our model outperforms its precursor. Neither of the two dictionaries used with word counts: Loughran and McDonald (2011) dictionary, or the Harvard Psychology dictionary, can match the performance of Linearized Phrase-Structure -model. Also, we find that Loughran and McDonald (2011) dictionary outperforms the Harvard Psychology dictionary, as Loughran and McDonald (2011) suggest. Hence, we can confirm that finance scholars should employ context dependent dictionaries when using word counts, as suggested by Loughran and McDonald (2011). Our results are based on a wide cross-section of different qualitative texts which attenuates the possibility that our results are driven by a sample specific pattern. Our benchmarking process is conducted in line with the latest norms of content analysis literature. Therefore, we conclude that our findings stand as unequivocal evidence in favor of the use of more sophisticated methodology in estimating investor sentiment. We argue that future research should also use more advanced natural language programming techniques when estimating investor sentiment; the methodology can substantially improve the accuracy of sentiment estimation.

After evaluating the performance of prior methodologies, and establishing that more complex methods are warranted, we move on to study the impact sentiment has on financial metrics. In light of our benchmarking, we would expect Linearized Phrase-Structure -model based sentiment variable to outperform word count based sentiment estimates when analyzing the relationship with financial metrics (more of this in section 6.4.3).

As we had expected, Linearized Phrase-Structure -model outperforms its rivals; however, to our surprise, word count based sentiment variables (both dictionaries) lack significance in all our specifications.[197] In other words, we do not find any cause to support a relationship between word count sentiment variables and our dependent variables. In fact, we hypothesize that prior literature's findings fall prey to data dredging concerns proposed by EMH advocates (e.g., Fama, 1991). In other words, the choice of event window, and the source of qualitative text, in combination with arbitrary time period for the study, and other methodological choices, are driving the results of extant literature. Indeed, in line with the thinking of Fama (1998), we argue that with the use of different methodology, the documented results would most likely disappear. We conclude by positing that prior literature has reported spurious regularities.

Another implication of our findings is the proof that Loughran and McDonald (2011) dictionary does not perform as a predictor of stock market performance outside the scope of 10-k reports. Loughran and McDonald suggest in their article that future research should study their dictionary in the context of different qualitative texts to determine the dictionary's performance outside 10-k reports. Our findings seem to indicate that a word count sentiment based on Loughran and McDonald dictionary fails to classify text outside the 10-k context for the purpose of predicting stock metrics. Nevertheless, we do find support that Loughran and McDonald dictionary does outperform Harvard Psychology dictionary in the task of classifying text.

[197] The few exceptions when word count variables are significant, do not pass a more closer inspection of significance. To elaborate, the coefficient signs of the variables are not consistent inside the 21 quarters; casting doubt on any inference drawn from them.

6.4.2 Impact on behavioral finance theory

Our findings do support some of the findings of prior behavioral finance literature. In particular, our results suggest that the distraction hypothesis suggested by Hirsleifer et al., (2009) does exist. Aggregate market news volume has a distracting effect on investors, resulting in abnormal profits and abnormal volume, as suggested by the hypothesis. Indeed, our results offer strong evidence in favor of limited attention theory that sets the ground for the distraction hypothesis.

Prior literature has hypothesized that firm specific news volume can proxy for attention grabbing firms, and that a firm specific news volume variable can be used to forecast returns and other financial metrics. Our findings indicate that firm specific news volume is not related to abnormal returns, and the relationship to abnormal volume may be limited to a short-term effect. However, it appears that firm specific news volume is indeed related to abnormal volatility – most likely due to the fact that noise traders are attracted towards attention grabbing stocks. Therefore, our results indicate that companies that actively attract news coverage, and attention, exhibit more volatility.

Qualitatively our findings seem to support our hypothesis that limited attention is driving the relationships between our dependent variables and our key independent variables. Indeed, both volume and return findings indicate underreaction which is in line with limited attention. On the other hand, underreaction can be explained by several different theories. For instance, anchoring can explain underreaction to new information. Another explanation, closely related to anchoring, is conservatism that can also explain underreaction when new information is not clearly related to the saliency of the model it is being used in[198]. However, as our multivariate specification is a holistic media model, we can examine the alternative theories with more precision. Indeed, aggregate market news volume, as discussed, has a statistically significant relationship with abnormal returns and volume. Aggregate market news volume relationship is explained by limited attention theory that predicts underreaction. Moreover, competing hypotheses for aggregate market news volume's impact are far more difficult to establish as aggregate market news volume per se has no other competing explanations. Therefore, as demonstrated by aggregate market news volume findings, we argue that limited attention is indeed a phenomenon affecting the decisions and actions of agents. We posit that limited

[198] In fact, such would be the case with qualitative information as valuation models employ point-estimates, meaning that qualitative estimates need to be transformed into quantitative estimates.

attention theory is the leading explanation behind the link between media factors and financial metrics.

As our results indicate underreaction, we can infer that our media variables appear to have new information content, as underreaction is the dominant effect in the case of new information content[199]. Consequently, we provide additional evidence to the field of research studying the informational content of qualitative texts, arguing that qualitative texts do indeed possess new information on fundamentals. Therefore, we conclude by positing that investors should not ignore the information content in qualitative texts, and that at the aggregate level, the market does incorporate the information in qualitative texts into asset prices.

6.4.3 Impact on link between sentiment and stock performance

To our disappointment, our method for estimating investor sentiment: the Linearized Phrase-Structure -model does not have a significant relationship with abnormal returns or abnormal volume. We recognize the fact that our study limits might explain the lack of significance. However, we also maintain that our methodology is superior to that of prior literature, and our study is – to our knowledge – the most robust study undertaken in the field. Therefore, we argue that in light of our findings, sentiment is not a driver of abnormal returns or volume.

Based on our findings, we cannot draw precise conclusions on market efficiency.[200] The reasoning is three-fold. First, our sentiment variables and firm specific news volume variable, with the majority of our controls, are unable to explain future variations in abnormal returns. Second, aggregate market news volume and momentum factors can be used to forecast abnormal returns, and therefore stand as a testimony against market efficiency. Third, we have not tested specific trading strategies to examine the economic feasibility of our findings in the scope of this study. Therefore, we cannot exclude the possibility that the market is efficient based on its economic definition of having no abnormal profits ex-post trading costs (e.g., Jensen, 1978). However, we can argue that the market is not efficient in allocating

[199] If qualitative texts would not have new information, the reaction would be characterized by an overreaction (e.g., Tetlock, 2007), as discussed in previous sections of this study, or by no reaction at all.

[200] We have not included abnormal volume in the efficiency discussion as EMH does not give clear predictions on the impact of efficiency and volume. Also, we have excluded volatility from the discussion as we view it to be subject to the critique of Schwert (1991) discussed in Section 2.

assets based on its stricter more theoretical definition of no relationship with abnormal returns.[201]

Even though Linearized Phrase-Structure -model variable is unable to explain variations in abnormal returns or volume, it can be used to forecast abnormal volatility variations. As such, Linearized Phrase-Structure -model can in fact act as a tool for profitable trading strategy that revolves around trading volatility. For instance, option pricing is based on the volatility of the underlying asset. In other words, if Linearized Phrase-Structure -model can be used to forecast future volatility of a firm's share, it can be used to forecast the price development of the option that is a derivative of the firm's share. Moreover, Linearized Phrase-Structure -model can have applications in forecasting index volatilities such as the VIX –index, and trading strategies leveraging the movements of such indices. We conclude that using Linearized Phrase-Structure -model to forecast volatility holds promise, and future research should study volatility trading strategies that leverage Linearized Phrase-Structure -model as a tool.

[201] Our findings offer strong support for underreaction as the dominant effect; therefore, questioning the '*even-split anomaly distribution'* efficiency argument of Fama (1991).

7 CONCLUSIONS

In this final section of our research, we start by summarizing our findings. We first conclude on sentiment methodology and then on the impact of various factors on stock price performance. We finish by discussing possible avenues for further research.

7.1 Sentiment methodology

With the growing demand for sentiment analysis tools in financial and economic applications, it is increasingly important to pay attention to the ability of the models to capture the domain-specific use of language. As observed in the previous studies on general sentiment analysis, it is well-known that models which work in one domain may not work well in another one. In particular, when considering the specialized vocabulary encountered in finance and economics, building a model requires a combination of expert information in the form of high-quality lexicons as well as more sophisticated learning algorithms which are better able to account for the contextual dependence of semantic orientations.

The commonly applied bag-of-words approaches impose several restrictions. In particular, by treating sentences as unordered collections of words, they fail to take syntactic information into account. We have obtained encouraging results with our Linearized Phrase-Structure - model in the financial domain, where we have made a substantial effort to take into account phrase-structure information and the way the semantic orientations of financial concepts are influenced by other parts of text. Our findings contribute to the methodology employed in investor sentiment estimation: we have shown that by using more sophisticated sentiment estimation methodology researchers can reach more accurate sentiment estimate, which can potentially further lead to better predictions for stock performance metrics, such as volatility.

Also, we have validated the assertion of Loughran and McDonald (2011) that states that researchers should employ context dependent dictionaries when using word count methodology in finance. Indeed, we posit that Loughran and McDonald’s (2011) dictionaries should be utilized as the preferred dictionary in future research that employs word counts.

7.2 Media and stock performance

Our second objective of testing the relation between media factors, both sentiment and others, with stock performance is far more ambitious. In fact, finding a strong relation here would be surprising, as this would potentially represent a market inefficiency that would likely be exploited rapidly in the financial markets. Our conclusions here relate to testing financial theories on the impact of media factors, the possibility that limited attention theory explains the impact of media factors, impact of qualitative information on financial metrics, and market efficiency.

First, we offer insights to the significance of the three major media factors (firm specific news volume, market news volume and sentiment, on financial metric) that have been studied in prior literature. We study the three major media factors simultaneously, and show that three factors affect markets in chorus, and are not mutually exclusive. Furthermore, we validate the findings of Hirsleifer et al. (2009) on the importance and impact of aggregate market news volume on financial metrics. However, we fail to find a statistically significant relationship between sentiment estimates and abnormal returns or volume. We hypothesize, that in light of our sentiment estimation methodology benchmarking, the sentiment findings of prior literature are described by sample specific patterns, and are in fact spurious in nature. Therefore, the theoretical link between sentiment estimates and financial metrics is not empirically validated. That being said, we find a statistically significant relationship between sentiment and future abnormal volatility. In addition, even though lacking in statistical significance, we find that qualitatively our findings indicate an underreaction to sentiment changes, in line with extant literature. As a result, we suggest that the limitations impacting our sentiment estimation methodology can result in such a significant measurement error in the sentiment estimate that our results fail to represent the true significance of the variable. Therefore, we conclude by suggesting that sentiment can have a relationship with financial metrics. Finally, we document that firm-specific news volume impacts abnormal volatility, and to some extent, abnormal volume. We do not find a relationship between abnormal returns and firm specific news volume. In fact, we hypothesize that the documented link between returns and firm specific news has in fact acted as a proxy for the relationship between aggregate market news volume and abnormal returns.

Second, based on our findings, we suggest that limited attention is in fact the underlying theory behind most of the findings related to the media factors. We find that aggregate market news volume findings correspond to the findings of prior literature. Aggregate market news volume is hypothesized, by Hirsleifer et al. (2009), to have a relationship between financial metrics due to limited attention – no other competing viable theoretical explanation for the relationship exists to our knowledge. Based on our findings, we can validate the aforementioned relationship. Therefore, our findings also support the theory of limited attention. Furthermore, as limited attention implies an underreaction, and qualitatively our findings indicate an underreaction pattern with all the media factors, we suggest that it is not a farfetched idea to suggest that limited attention is the theory explaining the underreaction pattern with other media factors. In fact, as limited attention is already a phenomenon in the market place, as is demonstrated by aggregate market news volume findings, it is plausible to assume that the impact is not isolated to aggregate market news volume. Therefore, we posit that limited attention is the theory explaining the relationship between media factors and financial metrics.

Third, we offer our findings as a contribution to the discussion relating to the impact qualitative texts have on financial metrics. We have consistently documented a qualitatively indicative underreaction pattern between sentiment and variations in financial metrics. As extant literature has documented that texts with overreaction are associated with framing effects, and texts with underreaction are associated with novel information content, we suggest that our findings indicate that qualitative texts have novel information content, and that investor reaction is dominantly led by information content over tone effects. However, as our sentiment findings lack statistical significance, our findings simply provide a qualitative indication of the aforementioned. Nevertheless, we find that aggregate market news volume impacts financial metrics, and that relationship is statistically significant. We infer that this clearly indicates that qualitative texts do matter, and they are considered by investors during decision making. Therefore, we argue that while aggregate market news volume sheds no light on the nature of the impact: novel information vis-à-vis tone, the finding does indicate that qualitative texts do play a role in financial markets. In tandem with our qualitative findings for sentiment estimates, we suggest that qualitative texts impact financial metrics, and retain some novel information content.

Fourth, in culmination, our findings provide significant evidence to the discussion relating to the state of market efficiency. We find that both momentum factors and aggregate market news volume explain future variations in abnormal returns. Indeed, our findings stand in stark contradiction to the theoretical definition of the efficient market hypothesis. Also, when applicable, our findings describe a pattern of underreaction that has been the dominant pattern relating to media factors. Therefore, the critique of Fama (1991) that anomalies are distributed with an even split, does not seem to hold with media factors. However, we have not tested the economic significance of our findings, and therefore cannot take a stand on whether or not our findings are in contradiction to the economic form of market efficiency (e.g., Jensen, 1978).

As our study has employed a wide cross-section of methodologies, and qualitative text sources, we argue that our findings are robust to methodology changes, and free from data dredging concerns. However, we recognize that the limitations of our study can materially impact our results, and the interpretations drawn from them. We urge future research to especially improve our methodology, and to focus on the areas we have listed for future research in order to improve our understanding of the importance of media factors in financial markets.

7.3 Avenues for future research

While we consider our findings to be robust in several dimensions, we feel that future research can improve our current understanding of the role that media factors play in financial markets. First, there are multiple ways we can improve on the way to measure sentiment. To elaborate on this, we continue by describing the features that an ideal, '*comprehensive sentiment model*', should possess to be able to detect sentiment in a human-like manner. Finally, we discuss the key research directions we see related to finance literature.

7.3.1 Sentiment methodology

While our sentiment estimation methodology is a significant improvement from prior simplistic word count methodology, our method has several caveats. One of the key difficulties in phrase-level sentiment analysis is the lack of deeper contextual information. Since each phrase has to be interpreted in isolation, a human reader would often prefer to see more context than one sentence in order to draw any conclusions. For instance, although acquisition events are commonly described in quite positive tone (e.g. “The acquisition of Sampo Bank makes strategic sense for DB"), it is quite uncertain whether they will add or destroy value. Therefore, human analysts tend to be generally skeptical about the positiveness of such news, and would require considerable amount of background information before wanting to judge statement positive or negative. In addition to the lack of prior knowledge on the companies, the LPS algorithms are unable to distinguish between advertising-like positiveness following from a company's own statements vs. independent reviews about the company. Also some events, such as nominations of new executives, are conventionally described in overly positive manner instead of reflecting the actual facts. An interesting problem is also the detection of roles or perspective in a sentence (i.e. from whose viewpoint do we interpret the sentence when multiple parties are involved). For example, a sentence may talk about a company getting money from another company after a legal case. Good news for someone may well be bad for the other.

Clearly, there is still a number of ways to improve the performance of the sentiment models reported in this paper. In addition to the enrichment of lexicon with weights for different concepts and events, an important direction for future research will be to examine how phrase-level models can be merged with content models. We recognize specifically the following development areas for future models:

- Developing a sentiment methodology able to identify the topic of a text, and to capture the relevance of that text. For example a study by Antweiler and Frank (2006), which uses an algorithm to identify news stories by their topic rather than their tone, does find some return predictability. As a further step, researchers could test for the impact that qualitative texts dealing with company's key products have on sentiment, possibly looking at media that does not even mention the company's name
- Creating a method that is able to assess the credibility of a text source in order to establish a proper weighing score for the text
- Studying alternative ways of aggregating sentiment besides the use of negative fractions: for instance, Das (2010) disagreement sentiment. Also, different "shades" of negativity, e.g. bad news of events that have happened vs. increase in risk could have a different impact on the stock performance
- Assembling dictionaries in different languages, and researching whether or not a sentiment constructed from them has any difference to sentiment constructed from source texts written in English
- Constructing a methodology able to differentiate between texts dealing with historical information, present information, and forecasted estimates. For example O'Hare (2009) points out that traditional media reports most likely news relating to a stock's past performance but only few statements about a company's future. However analysis about future performance of a stock should be the part that investors consider most

7.3.2 A comprehensive sentiment model

To foster future development efforts, we finish our discussion on sentiment methodology by describing what a future model for detecting sentiment could potentially look like. This comprehensive sentiment model would take all the media on a given day, analyze this information in the context of previously publicly known information, and give an estimate of

- What news this information says about the company’s fundamentals and how should this affect the company’s valuation (information)
- How this information may impact different investors and thereby the demand for the company’s stock (tone)

For this goal, we describe the parts of that a comprehensive model could consist of. In addition to illustrating the purpose and workings of each part of the model, we also give our early hypotheses on how this model could be implemented, and potential pitfalls for each part. The proposed model is described in Table 23.

Table 23: Proposed comprehensive sentiment model

Model part	Purpose	Example	Implementation	Pitfalls
Word detection	Detect polarized parts in text, possibly also with sentiment strength	Detects word "well" in a sentence "We believe that Samsung will not do well" and assigns labels to it, e.g. "reversal" and "positive"	Word lists for keywords of different categories	Use of flawed word lists
Pattern matching for polarized words	Translate a pattern of detected words to a sentiment, possibly with a score for strength	Based on training data, assigns label "negative to tags "reversal + positive"	Quasi compositional sequencing (SVM) with pattern compression	Training data that uses more knowledge than is available at this stage of the model
Credibility filtering	Assess the credibility of a sentence	Detects that the praising document is written by the company's sales department and reduces the strongly positive sentiment	Author recognition with entity detection, credibility assessment (entity-database from a Wikipedia-based ontology[202] + database of historical source sentiments)	Entity database lacking key entities, difficulty of assigning a credibility estimate to listed entities
Topic detection	Detect the key topic(s) in a sentence	Detects from "Samsung's new Galaxy is selling moderately well" that it relates to galaxy sales / smart phone sales	Topic recognition with a Wikipedia-based ontology	Hierarchy of topics, colloquial language
Impact assessment	Detect how much further impact will this media item have on the topic have when added on top of the existing information on the market	Detects whether a piece of information has been discussed before (Is it novel news/analysis?), confirms a previously uncertain information (Is this source more credible?), or if it reaches now a new audience (Did this information previously reach only people who read the specialized industry magazine?)	Database on information by topic for comparing novelty, impact factors for different sources based on reach, prestige etc.	Difficulty of novelty assessment, lack of impact factor data for news
Topic relevancy filtering	Assign a relevancy score per topic, highlighting the most important topics	Ranks e.g. the topic "mobile phones" higher than "laptops" for Nokia	A layered ontology of concepts: importance of concepts for all companies, for different industries and for individual companies, based on metrics such as word-of-mouth in media, company's own releases, and analyst reports	Difficulty of ranking concepts, hype of concepts vs. reality

[202] For more information on Wikipedia-based ontologies, see Malo et al., 2010

A sentiment model of this kind would be significantly closer to mimicking the key stages of an analyst's thoughts in assessing the impact of news on a company and could potentially form similar conclusions as financial analysts who study a company regularly. As for the moment, most studies have implemented only the first part of this model, which is likely to yield sentiment scores with significant biases. Though the proposed model includes some significant challenges in implementing, we find that models that go towards this direction could significantly improve the detected sentiment and make it more relevant in relation to stock performance.

7.3.3 Data and specifications

As with all studies, our study faced limitations in terms of data accessibility. Also, in light of our findings, there are areas that future study specifications should consider when conducting their research. We feel that the key areas of interest in the aforementioned areas are:

- Analyzing sentiment using social media text sources, such as, i.e., Twitter
- Studying the impact media factors have in Fringe markets (i.e. smaller countries such as Finland), and whether or not the impact differs from the findings of extant literature
- Using intra-day data when studying the impact media factors have on financial metrics
- Constructing trading strategies leveraging aggregate market news volume, and sentiment volatility forecasts, to measure economic feasibility of such trading schemes
- Studying the impact high firm specific news has on volatility, and whether or not that translate into a rise in cost of capital as time passes on

To conclude, we have casted some additional light on the big picture of how media factors and sentiment link to each other. Future research could complement these findings in particular by further improving on the research, and by tapping to further data sources.

8 REFERENCES

Scientific articles and working papers

Ajinkya B., Gift, M., 1984. Corporate managers' earnings forecasts and symmetrical adjustments of market expectations. Journal of Accounting Research 22, 425-444.

Akerlof, G., 1991. Procrastination and obedience. American Economic Review Papers & Proceedings 81, 1-19.

Alexander, S., 1961. Price movements in speculative markets: trends or random walks. Industrial Management Review 2, 7-26.

Antweiler, W., Frank, M., 2004. Is all that talk just noise? The information content of internet stock message boards. Journal of Finance 59, 1259-1294.

Antweiler, W., Frank, M., 2006. Do U.S. stock markets typically overreact to corporate news stories? Working Paper. University of British Columbia.

Arbit, H., Boldt, B., 1984. Efficient markets and the professional investor. Financial Analysts Journal 40, 22-34.

Ariel, R., 1987. A monthly effect in stock returns. Journal of Financial Econometrics 18, 161-174.

Ariel, R., 1990. High stock returns before holidays: existence and evidence on possible causes. Journal of Finance 45, 1611-1626.

Bachelier, L., 1900. Theorie de la speculation. Annales Scientifiques de l'Ecole Normale Superieure Ser 3, 21-86.

Baker, A., Valleettourangeau, P., Frank, R., Pan, M., 1993. Selective associations and causality judgements — presence of strong causal factor may reduce judgements of a weaker one. Journal of Experimental Psychology: Learnings, Memory, and Cognition 19, 414-432.

Ball, R., 1978. Anomalies in relationships between securities' yields and yield-surrogates. Journal of financial Economics 6, 103-1126.

Ball, R., 1996. The theory of stock market efficiency: Accomplishments and limitations. Journal of Financial Education 22, 1-13.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159-178.

Banz, R., 1981. The relationship between return and market value of common stocks. Journal of Financial Economics 9, 3-18.

Barber, B,. Odean, T., 2002. Online investors: Do the slow die first? Review of Financial Studies 15, 455-487.

Barber, B., Lyon, J., 1997. Detecting long-run abnormal stock returns: the empirical power and specification of test statistics. Journal of Financial Economics 43, 341-372.

Barber, B., Odean, T., 2000. Trading is hazardous to your wealth: the common stock performance of individual investors. Journal of Financial Economics 43, 341-372.

Barber, B., Odean, T., 2001. Boys will be boys: Gender, overconfidence, and common Stock investment. Quarterly Journal of Economics 141, 261-292.

Barber, B., Odean, T., 2008. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies 21, 785-818.

Barberis, N., Huang, M., 2001. Mental accounting, loss aversion, and individual stock returns. Journal of Finance 56, 1247-1292.

Barberis, N., Shleifer, A., 2003. Style investing. Journal of Financial Economics 68, 161-199.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. Journal of Financial Economics 49, 307-345.

Basu, S., 1977. Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. Journal of Finance 32, 663-682.

Basu, S., 1983. The relationship between earnings' yield, market value and the returns for NYSE common stocks: further evidence. Journal of Financial Economics 12, 129-156.

Baumeister, F., Bratslavsky, E., Finkenauer, C., Vohs, K., 2001. Bad is stronger than good. Review of General Psychology 5, 323-370.

Beechey, M., Gruend, D., Vickery, J., 2000. The efficient market hypothesis: a survey. Research Discussion Paper. Reserve Bank of Australia, Australia.

Bell, D., 1982. Regret in decision making under uncertainty. Operations Research 30, 961-981.

Benartzi, S., Thaler, R., 2001. Naïve diversification strategies in defined contribution savings plans. American Economic Review 91, 79-98.

Bernard, V., Thomas, J., 1989. Post-earnings-announcement drift: delayed price response or risk premium? Journal of Accounting Research, Supplement 27, 1-48.

Bernard, V., Thomas, J., 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics 13, 305-340.

Bhandari, L., 1988. Debt/equity ratio and expected common stock returns: empirical evidence. Journal of Finance 43, 507-528.

Bhattacharya, U., Galpin, R., Yu, X., 2009. The role of the media in the internet IPO bubble. Journal of Financial and Quantitative Analysis 44, 657-682.

Black, F., 1972. Capital market equilibrium with restricted borrowing. Journal of Business. 45, 444-54.

Bligh, M., Hess, G., 2007. The power of leading subtly: Alan Greenspan, rhetorical leadership, and monetary policy. The Leadership Quarterly 18, 87-104.

Bloomfield, R., 2002. The “incomplete revelation hypothesis” and financial reporting. Accounting Horizons 16, 233-243.

Boldt, B., Arbit, H., 1984. Efficient markets and the professional investor. Financial Analysts Journal 40, 22-34.

Brandt, M,. Kishore, R., Santa-Clara, P., Venkatachalam, M., 2008. Earnings announcements are full of surprises. Working Paper, Duke University.

Brav, A., 200. Inference in long-horizon event studies. Journal of Finance 55, 1979-2016.

Brav, A., Geczy, C., Gompers, P., 2000. Is the abnormal return following equity issuances anomalous? Journal of Financial Economics 56, 209-249.

Brav, A., Heaton, J., 2002. Competing theories of financial anomalies. Review of Financial Studies 15, 575-606.

Brennan, M., Chordia, T., Subrahmanyam, A., 1998. Alternative factor specifications, security characteristics and the cross-section of expected stock returns. Journal of financial Economics 49, 345-373.

Brenner, L., Koehler, D., Tversky, A., 1996. On the evaluation of one-sided evidence. Journal of Behavioral Decision Making 9, 59-70.

Brocas, I., Carillo, J., 2000. The value of information when preferences are dynamically inconsistent. European Economic Review 44, 1104-1115.

Brown, P., Kleidon, A., Marsh, T., 1983. New evidence on the nature of size related anomalies in stock prices. Journal of Financial Economics 12, 33-56.

Brown, R., 1828. A brief account of microscopical observations: made in the months of june, july and august, 1828, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Edinburg New Philosophical Journal 5, 358-371.

Brown, S., Pope, P., 1996. Post-earnings announcement drift? Working paper. New York University, New York.

Buehler, R., Griffin, D., Ross, M., 1994. Exploring the planning fallacy: why people underestimate their task completion time. Journal of Personality and Social Psychology 67, 366-381.

Busemeyer, J., Myung, J., McDaniel, M., 1993. Cue competition effects — empirical tests of adaptive network learning models. Psychological Science 4, 190-195.

Bushee, B., Core, J., Guay, W., Wee, J., 2010. The role of the business press as an information intermediary. Journal of Accounting 48, 1-19.

Cahan, R., Luo, Y., Alvarez, M., Jussa, J., Chen, Z., Wang, S., 2011. Signal processing Quant 2.0 – Harnessing the power of the web in quantitative investing. Deutsche Bank Global Markets Research.

Caillaud, B., Jullien B., 2000. Modeling time-inconsistent preferences. European Economic Review 44, 1116-1124.

Camerer C., Weber, M., 1992. Recent developments in modeling preferences: Uncertainty and ambiguity. Journal of Risk and Uncertainty 5, 325-370.

Camerer, C., 1998. Bounded rationality in individual decision making. Experimental Economics 1, 163-183.

Camerer, C., Hogarth, R., 1999. The effects of financial incentives in experiments: a review and capital-labor production framework. Journal of Risk and Uncertainty 19, 7-42.

Campbell, J., Shiller, R., 1988. Stock prices, earnings, and expected dividends. Journal of Finance 43, 661-676.

Campbell, J., Shiller, R., 1988. Stock prices, earnings, and expected dividends. Journal of Finance 43, 661-676.

Campbell, J., Shiller, R., 1998. Valuation ratios and the long-run stock market outlook. Journal of Portfolio Management 24, 11-26.

Chaiken, S., 1987. The heuristic model of persuasion. In Social Influence: The Ontario Symposium 5, 3-40.

Chan, W., 2003. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics 70, 223-260.

Charest, G., 1978. Dividend information, stock returns and market efficiency-II. Journal of Financial Economics 6, 463-465.

Chen, C., Lee, C., Lin, W., Yen, G., 2001. On the Chinese lunar New Year effect in six Asian stock markets: an empirical analysis. Review of Pacific Basin Financial Markets and Policies 4, 463-478.

Cherry, E., 1953. Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustical Society of America 25, 975-979.

Chew, S., 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. Econometrica 51, 1065-1092.

Chew, S., 1989. Axiomatic utility theories with the betweenness property. Annals of Operations Research 19, 273-298.

Chopra, N., Lakonishok, J., Ritter, J., 1992. Measuring abnormal performance: do stocks overreact? Journal of Financial Economics 31, 235-268.

Chung, C., Pennebaker, J., 2007. The psychological functions of function words. Social Communication, 343-359.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. Journal of Finance 63, 1977-2011.

Constantinides, G., 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. Journal of Business 55, 253-267.

Cowles, A., 1933. Can stock market forecasters forecast? Econometrica 1, 309-324.

Cowles, A., 1944. Stock market forecasting. Econometrica 12, 206-214.

Cowles, A., 1960. A revision of previous conclusions regarding stock price behavior. Econometrica 28, 909-915.

Cowles, A., Jones, H., 1937. Some a posteriori probabilities in stock market action. Econometrica 5, 280-294.

Cross, F., 1973. The behavior of stock prices on Fridays and Mondays. Financial analysts Journal 29, 67-69.

Cutler, D., Poterba, J., Summers, L., 1989. What moves stock prices? Journal of Portfolio Management 15, 4-12.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor psychology and security market under- and overreactions. Journal of Finance 53, 1839-1885.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 2001. Overconfidence, arbitrage and equilibrium asset pricing. Journal of Finance 56, 921-965.

Daniel, K., Hirshleifer, D., Teoh, S., 2002. Investor psychology in capital markets: evidence and policy implications. Journal of Monetary Economics 49, 139-209.

Daniel, K., Titman, S., 1997. Evidence on the characteristics of cross-sectional variation in common stock returns. Journal of Finance 52, 1-33.

Das, S., Chen, M., 2006. Yahoo! for amazon: Sentiment extraction from small talk on the web. Working Paper, Santa Clara University.

Davis, A., Piger, J., Sedor, L., 2008. Beyond the numbers: Managers' use of optimistic and pessimistic tone in earnings press releases. Working Paper. Federal Reserve Bank of St. Louis.

De Long, J., Shleifer, A., Summers, L., Waldmann, R., 1990. Noise trader risk in financial markets. Journal of Political Economy 98, 703-738.

DeBondt, W., Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793-805.

DeBondt, W., Thaler, R., 1987. Further evidence on investor overreaction and stock market seasonality. Journal of Finance 42, 557-581.

Dekel, E., 1986. An axiomatic characterization of preferences under uncertainty: weakening the independence axiom. Journal of Economic Theory 40, 304-318.

Demers, E., Vega, C., 2010. Soft information in earnings announcements: news or noise? Working Paper. INSEAD.

Dimson, E., Mussavian, M., 1998. A brief history of market efficiency. European Financial Management 4, 91-103.

Dovring, K., 1954. Quantitative semantics in 18th century Sweden. Public Opinion Quarterly 18, 389-394.

Dye, R., Sridhar, S., 2004. Reliability-relevance trade-offs and the efficiency of aggregation. Journal of Accounting Research 42, 51-88.

Einstein, A., 1905. Über die von der molekularkinetischen Theorie der warmen geforderten Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik 322, 549–560.

Ellsberg, D., 1961. Risk, ambiguity, and the savage axioms. Quarterly Journal of Economics 75, 643-669.

Engelberg, J., 2008. Costly information processing: Evidence from earnings announcements. Working Paper. Northwestern University.

Fama, E., 1965. The behavior of stock market prices. Journal of Business 38, 34-105.

Fama, E., 1970. Efficient capital markets: a review of theory and empirical work. Journal of Finance 25, 383-417.

Fama, E., 1991. Efficient capital markets: II. Journal of Finance 46, 1575-1617.

Fama, E., 1998. Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics 49, 283-306.

Fama, E., Fisher, L., Jensen, M., Roll, R., 1969. The adjustment of stock prices to new information. International Economic Review 10, 1-21.

Fama, E., French, K., 1988a. Permanent and temporary components of stock prices. Journal of Political Economy 96, 246-273.

Fama, E., French, K., 1988b. Dividend yields and expected stock returns. Journal of Financial Economics 22, 3-22.

Fama, E., French, K., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25, 23-49.

Fama, E., French, K., 1992. The cross-section of expected stock returns. Journal of Finance 47, 427-465.

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Finance 333, 3-56.

Fama, E., French, K., 1996. Multifactor explanations of asset pricing anomalies. Journal of Finance 51, 55-84.

Fama, E., French, K., 1997. Industry costs of equity. Journal of Financial Economics 43, 153-193.

Fama, E., French, K., 2006. Profitability, investment, and average returns. Journal of Financial Economics 82, 491-518.

Fama, E., MacBeth, J., 1973. Risk, return and equilibrium: Empirical tests. Journal of Political Economy 81, 607-636.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. Journal of Finance 64, 2023-2052.

Farmer, J., Lo, A., 1999. Frontiers of finance: Evolution and efficient markets. Proceedings of the National Academy of Sciences of the United States of America 96, 991-992.

Fischer, C., 2001. Read this paper later: Procrastination with time-inconsistent preferences. Journal of Economic Behavior and Organization 46, 249-269.

Fischhoff, B., Slovic, P., Lichtenstein S,. 1977. Knowing with certainty: the appropriateness of extreme confidence. Journal of Experimental Psychology 3, 552-564.

Fisher, K., Statman, M., 2000. Cognitive biases in market forecasts. Journal of Portfolio Management 27, 72-81.

Foster, G., Shevlin, T., 1984. Earnings releases, anomalies, and the behavior of securities returns. The Accounting Review 59, 574-603.

Fox, C., Tversky, A., 1995. Ambiguity aversion and comparative ignorance. Quarterly Journal of Economics 110, 585-603.

Frankfurter, G., McGoun, E., 2002. Resistance is futile: The assimilation of behavioral finance. Journal of Economic Behavior and Organization 48, 375-389.

Frazzini, A., 2006. The disposition effect and underreaction to news. Journal of Finance 61, 2017-2046.

Frederick, S., Loewenstein, G., O’Donoghue, T., 2002. Time discounting and time preference: a critical review. Journal of Economic Literature 40, 351-401.

Freeman, R., Tse, S 1989. The multi-period information content of accounting earnings: confirmations and contradictions of previous earnings reports. Journal of Accounting Research, Supplement 27, 49-79.

French, K., 1980. Stock returns and the weekend effect. Journal of Financial Economics 8, 55-69.

French, K., Poterba, J., 1991. Investor diversification and international equity markets. American Economic Review 81, 222-226.

Friedman, M., 1953. The Case for Flexible Exchange Rates. Essays in Positive Economics. University of Chicago Press, Chicago.

Froot, K., Dabora, E., 1999. How are stock prices affected by the location of trade? Journal of Financial Economics 53, 189-216.

Gibbons, M., Hess, P., 1981. Day of the week effects and asset returns. Journal of Business 54, 3-27.

Gilovich, T., Vallonen, R., Tversky, A., 1985. The hot hand in basketball: on the misperception of random sequences. Cognitive Psychology 17, 295-314.

Glassman, C., 2005. Remarks at the plain language association international’s fifth international conference.

Griffin, D., Tversky, A., 1992. The weighing of evidence and the determinants of confidence. Cognitive Psychology 24, 411-435.

Griffin, P., 2003. Got information? Investor response to form 10-k and form 10-q EDGAR filings. Review of Accounting Studies 8, 433-460.

Grinblatt, M., Keloharju, M., 2001. How distance, language, and culture influence stockholdings and trades. Journal of Finance 56, 1053-1073.

Grossman, S., Shiller, R., 1981. The determinants of the variability of stock market prices. American Economic Review 71, 222-227.

Grossman, S., Stiglitz, J., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393-408.

Grullon, G., Michaely, R., 2004. The information content of share repurchase programs. Journal of Finance 59, 651-680.

Gul, F., 1991. A theory of disappointment in decision making under uncertainty. Econometrica 59, 667-686.

Hansen, L., Jagannathan, R., 1991. Implications of security market data for models of dynamic economies. Journal of Political Economy 99, 252-262.

Harvey, J., 1998. Herustic judgement theory. Journal of Economic Issues 32, 47-64.

Hassel, J., Jennings, R., 1986. Relative forecast accuracy and the timing of earnings forecast announcements. The Accounting Review 61, 58-75.

Haug, M., Hirschey, M., 2006. The January effect. Financial Analysts Journal 62, 78-88.

Haugen, R., Jorion, P., 1996. The January effect: Still there after all these years. Financial Analysts Journal 52, 27-31.

Heath, C., Tversky, A., 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of Risk and Uncertainty 4, 5-28.

Heaton, J., 2002. Managerial optimism and corporate finance. Financial Management 31, 33-45.

Henry, E,. 2008. Are investors influenced by the way earnings press releases are written? Journal of Business Communication 45, 363-407.

Hirshleifer, D., 2001. Investor psychology and asset pricing. Journal of Finance 56, 1533-1597.

Hirshleifer, D., Hong, S., 2002. Limited attention, information disclosure and financial reporting. Journal of Accounting and Economics 36, 337-386.

Hirshleifer, D., Lim, S., Teoh, S., 2009. Driven to distraction: Extraneous events and underreaction to earnings news. Journal of Finance 64, 2289-2325.

Hirsleifer, D., Teoh, S., 2003. Limited attention, financial reporting and disclosure. Journal of Accounting and Economics 36, 337-386.

Hirsleifer, D., Teoh, S., 2005. Limited investor attention and stock market misreactions to accounting information. Working Paper, Ohio State University.

Ho, T., Michaely, R., 1988. Information quality and market efficiency. Journal of Financial and Quantitative Analysis 5, 357-386.

Hong, H,. Torous, W., Valkanow, R., 2007. Do industries lead stock markets? Journal of Financial Economics 83, 367-396.

Hong, H., Stein, J., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance 54, 2143-2184.

Hong, H., Stein, J., 2003. Differences of opinion, short-sales constraints, and market crashes. Review of Financial Studies 16, 487-525.

Hou, K., 2007. Industry information diffusion and the lead-lag effect in stock returns. Review of Financial Studies 20, 1113-1138.

Hsueh, P.-Y., Melville, P., & Sindhwani, V., 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (pp. 27-35). Stroudsburg, PA, USA: Association for Computational Linguistics.

Huberman, G,. 2001. Familiarity breeds investment. Review of Financial Studies 14, 659-680.

Huberman, G., Regev, T., 2001. Contagious speculation and a cure for cancer. Journal of Finance 56, 387-396.

Ikenberry, D., Lakonishok, J., Vermaelen, T., 1995. Market underreaction to open market share repurchases, Journal of Financial Economics 39, 181-208.

Ikenberry, D., Lakonishok, J., Vermaelen, T., 2000. Stock repurchases in Canada. Journal of Finance 55, 2373–2397.

Jegadeesh, N., 2000. Long-run performance of seasoned equity offerings: Benchmark errors and biases in expectations. Financial Management 29, 5-30.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48, 65-91.

Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: an evaluation of alternative explanations. Journal of Finance 56, 699-720.

Jensen, M., 1978. Some anomalous evidence regarding market efficiency. Journal of Financial Economics 6, 95-102.

Jones, C., Lamont, O., 2002. Short sale constraints and stock returns. Journal of Financial Economics 66, 207-239.

Kahneman, D., 2003. Maps of bounded rationality: Psychology for behavioral economics. The American Economic Review 93, 1449-1475.

Kahneman, D., Slovic, P., Tversky, A., 1982. Judgement under uncertainty: Heuristics and biases. Science 185, 1126- 1131.

Kahneman, D., Tversky, A., 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 1124-1131.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263-292.

Kahneman, D., Tversky, A., 2000. Choices, values and frames. American Psychologist 39, 341-350.

Keim, D., 1983. Size-related anomalies and stock return seasonality: further empirical evidence. Journal of Financial Economics 12, 13-32.

Keim, D., Stambaugh, R., 1984. A further investigation of the weekend effect in stock returns. Journal of Finance 39, 819-835.

Keynes, J., 1923. Some aspects of commodity markets. Manchester Guardian Commercial: European Reconstruction Series, 784-786.

King, R., Pownall G., Waymire, G., 1990. Expectations adjustment via timely management of forecasts: Review, synthesis, and suggestions for future research. Journal of Accounting Literature 9, 114-144.

Klibanoff, P., Lamont, O., Wizman, T., 1999. Investor reaction to salient news in closed-end country funds. Journal of Finance 53, 673-699.

Kothari, S., Li, X,. Short, J,. 2008. The effect of disclosure by management, analysts, and financial press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. Working Paper, MIT.

Krishna, V., Morgan, J., 2004. The art of conversation: Eliciting information from experts through multi-stage communication. Journal of Economic Theory 117, 147-179.

Kruglanski, A., Thompson, E., 1999. Persuasion by a single route: a view from the unimodel. Psychological Inquiry 19, 83-109.

Kruschke J., Johansen, M., 1999. A model of probabilistic category learning. Journal of Experimental Psychology: Learnings, Memory, and Cognition 25, 1083-1119.

Laibson, D., 1997. Golden eggs and hyperbolic discounting. Quarterly Journal of Economics 112, 443-477.

Lakonishok, J., Levi, M., 1982. Weekend effects on stock returns: a note. Journal of Finance 37, 883-889.

Lakonishok, J., Smidt, S., 1988. Are seasonal anomalies real? A ninety-year perspective. Review of Financial Studies 1, 403-425.

Lakonishok, J., Vermaelen, T., 1990. Anomalous price behavior around repurchase tender offers. Journal of Finance 45, 455-477.

Lamont, O., Thaler, R., 2003. Can the market add and subtract? Mispricing in tech stock carve-outs. Journal of Political Economy 111, 227-268.

Langlois, R., 2003. Cognitive comparative advantage and the organisation of work: Lessons from Herbert Simon’s vision of the future. Journal of Economics Psychology 24, 167-187.

Lee, C., Shleifer A., Thaler, R., 1991. Investor sentiment and the closed-end fund puzzle. Journal of Finance 46, 75-110.

Lee, C., Yen, G., 2008. Efficient market hypothesis (EMH): past, present and future. Review of Pacific Basin Financial Markets and Policies 11, 305-329.

Lee, C., Yen, G., Chang, C., 1993. Informational efficiency of capital market revisited: anomalous evidence from a refined test. Advances in Quantitative Finance and Accounting 2, 39-65.

LeRoy, S., 1989. Efficient capital markets and martingales. Journal of Economic Literature 27, 1583-1621.

LeRoy, S., Porter, D., 1981. The present-value relation: Tests based on implied variance bounds. Econometrica 49, 555-574

Lewellen, J., 2002. Momentum and autocorrelation in stock returns. Review of Financial Studies 15, 533-564.

Lewis, K., 1999. Trying to explain home bias in equities and consumption. Journal of Economic Literature 37, 571-608.

Li, F., 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Working Paper. University of Michigan.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221-247.

Li, F., 2009. The determinants and information content of the forward-looking statements in corporate filings - a naïve Bayesian machine learning approach. Working Paper, University of Michigan.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13-37.

Lo, A., MacKinlay, A., 1990. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431-467.

Lo, A., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: Computational algorithms, statistical inference and empirical implementation. Journal of Finance 55, 1705-1765.

Long, J., 1978. The market valuation of cash dividends: a case to consider. Journal of Financial Economics 6, 235-264.

Loomes, G., Sugden, R., 1982. Regret theory: an alternative theory of rational choice under uncertainty. The Economic Journal 92, 805-824.

Lord, C., Ross, L., Lepper, M., 1979. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology 37, 2098-2109.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of finance 66, 35-65.

Loughran, T., Ritter, J., 1995. The new issues puzzle. Journal of Finance 50, 23-51.

Loukusa, P., 2011. Media coverage and the cross section of stock returns: Evidence from UK markets. Master's Thesis. Aalto University School of Economics, Helsinki.

Lowenstein, G., 2000. Emotions in economic theory and economic behavior. American Economic Review 65, 426-432.

Maks, I., and Vossen, P., 2010. Annotation Scheme and Gold Standard for Dutch Language Resources and Evaluation (LREC-10). European Language Resources Association (ELRA).

Malkiel, B., 2003. The efficient market hypothesis and its critics. Journal of Economics Perspectives 17, 52-82.

Malmendier, U., Tate, G., 2005. CEO overconfidence and corporate investment. Journal of Finance 60, 2661-2700.

Malo, P., Siitari, P., Ahlgren, O., Wallenius, J. and Korhonen, P., 2010. Semantic Content Filtering with Wikipedia and Ontologies. In Proceedings of IEEE International Conference on Data Mining Workshops, 2010, pp. 518--526.

Malo, P., Sinha, A., Takala, P., Ahlgren, O. and Lappalainen, I., 2013. Capturing sentiments in financial news: Towards knowledge-driven tree kernels. Working paper, submitted for The International Conference on Knowledge Discovery and Data Mining (KDD).

Malo, P., Sinha, A., Takala, P., Korhonen, P. and Wallenius, J., 2013. Good debt or bad debt: Detecting semantic orientations in economic texts. Working paper, submitted for Journal of the American Society for Information Science and Technology (JASIST).

Markowitz, H., 1952. The utility of wealth. Journal of Political Economy 60, 151-158.

Mercer, M., 2004. How do investors assess the credibility of management disclosures? Accounting Horizons 18, 185-196.

Merton, R., 1985. On the current state of stock market rationality hypothesis. Working paper, 1717-85. Sloan School of Management, Boston.

Michaely, R., Thaler, R., Womack, K., 1995. Price reactions to dividend initiations and omissions. Journal of Finance 38, 1597-1606.

Miller, E., 1977. Risk, uncertainty and divergence of opinion. Journal of Finance 32, 1151-1168.

Mitchell, M., Mulherin, J., 1994. The impact of public information on the stock market. Journal of Finance 49, 923-950.

Mitchell, M., Stafford, E., 1997. Managerial decisions and long-term stock price performance. Working paper. Graduate School of Business, University of Chicago.

Mitra, L., & Mitra, G., 2010. Applications of news analytics in finance: A review (Tech. Rep.). optirisk-systems.com/papers/Opt0014.pdf: OptiRisk Systems.

Modigliani, F., Cohn, R., 1979. Inflation and the stock market. Financial Analysts Journal 35, 24-44.

Moilanen, K., Pulman, S., & Zhang, Y., 2010. Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression. In Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (wassa 2010) at the 19th European conference on artificial intelligence (ecai 2010) (pp. 36{43). Lisbon, Portugal.

Moray, N., 1959. Attention in dichotic listening: Affective cues and the influence of instructions. Quarterly Journal of Experimental Psychology 11, 56-60.

Morris, M., Sheldon, O., Ames, D., Young, M., 2005. Metaphor in stock market commentary: Consequences and preconditions of agentic descriptions of price trends. Working Paper, Columbia University.

Mullainathan, S., 2001. Thinking through categories. Working Paper. MIT Press, Massachusetts.

Mullainathan, S., Thaler, R., 2000. Behavioral economics. Working Paper. NBER.

Newey, W., West, K., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703-708.

Nichols, N., 1993. Efficient? Chaotic? What's the new finance? Harvard Business Review 71, 50-56.

Nicholson, S., 1960. Price-earnings ratios. Financial Analysts Journal 16, 43-50.

Niederhoffer, V., Osborne, M., 1966. Market making and reversal on the stock exchange. Journal of the American Statistical Association 61, 897-916.

O'Donoghue, T., Rabin, M., 1999a. Incentives for procrastinators. Quarterly Journal of Economics 114, 769-816.

O'Donoghue, T., Rabin, M., 1999b. Doing it now or later. American Economic Review 89, 103-124.

O'Donoghue, T., Rabin, M., 2001. Choice and procrastination. Quarterly Journal of Economics 116, 121-160.

Odean, T., 1998a. Are investors reluctant to realize their losses? Journal of Finance 53, 1775-1798.

Odean, T., 1998b. Volume, volatility, price, and profit when all traders are above average. Journal of Finance 53, 1887-1934.

Odean, T., 2000. Do investors trade too much? American Economic Review 89, 1279-1298.

Pashler, H., & Johnston, J. C., 1998. Attentional limitations in dual-task performance. In H. Pashler (Ed.), Attention, Psychology Press/Erlbaum (Uk) Taylor & Francis, Hove, England, 155-189 .

Petersen, M,. 2004. Information: Hard and soft. Working Paper, Northwestern University.

Petersen, M., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. Review of Financial Studies 22, 435-480.

Petty, R., Cacioppo, J., 1986. Communication and persuasion: Central and peripheral routes to attitude change. Springer, New York.

Plumlee, M., 2003. The effect of information complexity on analysts' use of that information. Accounting Review 78, 275-296.

Pontiff, J., 1996. Costly arbitrage: Evidence from closed-end funds. Quarterly Journal of Economics 111, 1135-1151.

Poterba, J., Summers, L., 1988. Mean reversion in stock returns: Evidence and implications. Journal of Financial Economics 22, 27-59.

Quiggin, J., 1982. A theory of anticipated utility. Journal of Economic Behavior and Organization 3, 323-343.

Rabin, M., 1998. Psychology and economics. Journal of Economics Literature 36, 11-46.

Rabin, M., 2002. Inference by believers in the law of small numbers. Quarterly Journal of Economics 117, 775-816.

Rabin, M., Schrag, J., 1999. First impressions matter: a model of confirmatory bias. Quarterly Journal of Economics 114, 37-82.

Read, D., Lowenstein, G., Rabin, M., 1999. Choice bracketing. Journal of Risk and Uncertainty 19, 171-197.

Reinganum, M., 1981. Misspecification of capital asset pricing: Empirical anomalies based on earnings yields and market values. Journal of Financial Economics 9, 19-46.

Reingaum, M., 1981. The anomalous stock market behavior of small firms in January: Empirical tests for tax-loss selling effects. Journal of Financial Economics 12, 89-104.

Rendleman, R., Jones, C., Latané, H., 1987. Further insight into the standardized unexpected earnings anomaly: Size and serial correlation effects. Financial Review 22, 131-144.

Richards, A., 1997. Winner-loser reversals in national stock market indices: Can they be explained? Journal of Finance 52, 2129-2144.

Ritter, J., 1988. The buying and selling behavior of individual investors at the turn of the year. Journal of Finance 43, 701-719.

Ritter, J., 2003. Behavioral finance. Pacific-Basin Finance Journal 11, 429-437.

Ritter, J., Warr, R., 2002. The decline of inflation and the bull market of 1982 to 1997. Journal of Financial Quantitative Analysis 37, 29-61.

Roberts, H., 1967. Statistical versus clinical prediction of the stock market. Unpublished manuscript. University of Chicago, Chicago.

Rogalski, R., 1984. New findings regarding day-of-the-week returns over trading and non-trading periods: a note. Journal of Finance 39, 1603-1614.

Roll, R., 1986. The hubris hypothesis of corporate takeovers. Journal of Business 59, 197-216.

Roll, R., 1988. R-squared. Journal of Finance 43, 541-566.

Roll, R., Shiller, R., 1992. Comments: Symposium on volatility in U.S. and Japanese stock markets. Journal of Applied Corporate Finance 5, 25-29.

Romer, P., 2000. Thinking and feeling. American Economic Review Papers and Proceedings 90, 439-443.

Rouwenhorst, K., 1998. International momentum strategies. Journal of Finance 46, 3-27.

Rouwenhorst, K., 1999. Local return factors and turnover in emerging stock markets. Journal of Finance 54, 1439-1464.

Rozeff, M., 1984. Dividend yields are equity risk premiums. Journal of Portfolio Management 11, 68-75.

Scholes, M., 1969. A test of the competitive hypothesis: the market for new issues and secondary offerings. Unpublished PhD thesis. University of Chicago, Chicago.

Schwert, G., 1983. Size and stock returns, and other empirical regularities. Journal of Financial Economics 12, 3-12.

Schwert, G.,1991. Review of market volatility by R. Shiller: much ado about…very little. Journal of Portfolio Management 17,74-78.

Segal, U., 1987. Some remarks on Quiggin's anticipated utility. Journal of Economic Behavior and Organization 8, 145-154.

Segal, U., 1989. Anticipated utility: a measure representation approach. Annals of Operations Research 19, 359-373.

Sewell, M., 2011. History of the efficient market hypothesis. Research Note RN/11/04, University College London, London.

Shafir, E., Diamond, P., Tversky, A., 1997. “Money illusion”. Quarterly Journal of Economics 112, 341-374.

Sharpe, W., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance 19, 425-442.

Shefrin, H., Statman, .M., 1994. Behavioral capital asset pricing theory. Journal of Financial and Quantitative Analysis 29, 323-349.

Shefrin, H., Statman, .M., 2000. Behavioral portfolio theory. Journal of Financial and Quantitative Analysis 35, 127-151.

Shefrin, H., Statman, M,. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. Journal of Finance 40, 777-790.

Shefrin, H., Statman, M., 1984. Explaining investor preference for cash dividends. Journal of Financial Economics 13, 253-282.

Shiller, R., 1981. Do stock prices move too much to be justified by subsequent changes in dividends? American Economic Review 71, 421-436.

Shiller, R., 1982. Consumption, asset markets and macroeconomic fluctuations. Carnegie-Rochester Conference Series on Public Policy 17, 203-238.

Shiller, R., 1984. Stock prices and social dynamics. Brookings papers on economic activity 2, 457-510.

Shiller, R., 2000b. Measuring bubble expectations and investor confidence. Journal of Psychology and Financial Markets 1, 49-60.

Shiller, R., 2003. From efficient markets theory to behavioral finance. Journal of Economic Perspectives 17, 83-104.

Shleifer, A., Summer, L., 1990. The noise trader approach to finance. Journal of Economic Perspectives 4, 19-33.

Shleifer, A., Vishny, R., 1997. The limits of arbitrage. Journal of Finance 52, 35-55.

Simon, H., 1986. Rationality in psychology and economics. Journal of Business 59, 209-224.

Simons, D., Chabris, C., 1999. Gorillas in our midst: Sustained in attentional blindness for dynamic events. Perception 28, 1059-1074.

Simons, D., Levin, D., 1997. Change blindness. Trens in Cognitive Sciences 1, 261-267.

Skinner, D., 1994. Why firms voluntarily disclose bad news. Journal of Accounting Research 32, 38-60.

Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Accounting Review 71, 289-315.

Smirlock, M., Starks, L., 1985. Day-of-the-week and intraday effects in stock returns. Journal of Financial Economics 17, 197-210.

Smith, V., Suchanek, G,. Arlington, W., 1988. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. Econometrica 56, 1119-1153.

Smola, A.J., and Scholkopf, B (1998). “A Tutorial on Support Vector Regression," NeuroCOLT2 Technical Report, ESPIRIT Working Group in Neural and Computational Learning II.

Somasundaran, S., Ruppenhofer, J., & Wiebe, J., 2007. Detecting Arguing and Sentiment in Meetings. In Proceedings of the SIGdial Workshop on Discourse and Dialogue.

Spiess, K., Affleck-Graves, J., 1995. The long-run performance following seasoned equity issues. Journal of Financial Economics 54, 45-73.

Starmer, C., 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. Journal of Economic Literature 37, 332-382.

Starmer, C., Sugden, R., 1989. Violations of the independence axiom in common ratio problems: an experimental test of some competing hypotheses. Annals of Operational Research 19, 79-102.

Statman, M., 1999. Behavioral finance: Past battles and future engagements. Financial Analysts Journal 55, 12-17.

Stein, J., 2002. Information production and capital allocation: Decentralized versus hierarchical firms. Journal of Finance 57, 1891-1921.

Stracca, L., 2004. Behavioral finance and asset prices: Where do we stand? Journal of Economic Psychology 25, 373-405.

Stroop, J., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 28, 643-662.

Subramanian, R., Insley, R., Blackwell, R., 1993. Performance and readability: a comparison of annual reports of profitable and unprofitable corporations. Journal of Business Communication 30, 49-61.

Taylor, S., 1982. Tests of random walk hypothesis against a price trend hypothesis. Journal of Financial and Quantitative Analysis 17, 37-61.

Teoh, S., Welch, I., Wong, T., 1998. Earnings management and the underperformance of seasoned equity offerings. Journal of Financial Economics 50, 63-69.

Tetlock, P., 2007. Giving content to investor sentiment: the role of media in the stock market. Journal of Finance 62, 1139-1168.

Tetlock, P., Saar-Tsechansky, M., Macskassy, 2008. More than words: Quantifying language to measure firms’ fundamentals. Journal of Finance 63, 1437-1467.

Thaler, R., 1980. Toward a positive theory of consumer choice. Journal of Economic Behavior and Organization 1, 39-60.

Thaler, R., 1999. The end of behavioral finance. Financial Analysts Journal 55, 12-17.

Thiele, T., 1880. Om anvendelse af mindste kvadraters methode i nogle tilfælde, hvor en komplikation af visse slags uensartede tilfældige fejlkilder giver fejlene en ‘systematisk’ karakter. Vidensk. Selsk. Skr. 5. Rk., naturvid. og mat. Afd. 12, 381–408.

Thomson, R., 1978. The information content of discounts and premiums on closed-end fund shares. Journal of Financial Economics 6, 151-186.

Trautmann, B., Hamilton, G., 2003. Informal corporate disclosure under federal securities law: Press releases, analyst calls, and other communications. Chicago, IL.

Tversky, A., Kahneman D., 1986. Rational choice and the framing of decisions. Journal of Business 59, 251-278.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297-323.

Tversky, A., Thaler, R., 1990. Anomalies: Preferences reversals. Journal of Economic Perspectives 4, 201-211.

Vapnik, V (1995). The Nature of Statistical Learning Theory, Springer-Verlag, New York.

Vapnik, V, and A. Lerner (1963). Pattern Recognition using Generalized Portrait Method. Automation and Remote Control, v24.

Vapnik, V. and Chervonenkis (1964). \On the Uniform Convergence of Relative Frequencies of Events to their Probabilities," Theory of Probability and its Applications, v16(2), 264-280.

Verrecchia, R., 2001. Essays on disclosure. Journal of Accounting and Economics 32, 97-180.

Weinstein N., 1980. Unrealistic optimism about future life events. Journal of Personality and Social Psychology 39, 806-820.

West, K., 1988. Dividend innovations and stock price volatility. Econometrica 56, 37-61.

White, H., 1984. A heteroskedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity. Econometrica 48, 817–38.

Wiebe, J., Wilson, T., & Cardie, C., 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39 , 165-210.

Williamson, P., 1997. Learning and bounded rationality. Journal of Economic Surveys 11, 221-230.

Working, H., 1960. Note on the correlation of first differences of averages in a random chain. Econometrica 28, 916-918.

Wurgler, J., Zhuravskaya K., 2002. Does arbitrage flatten demand curves for stocks? Journal of Business 75, 583-608.

Yaari, M., 1987. The dual theory of choice under risk. Econometrica 55, 95-115.

Books

Alpert, M., Raiffa, H., 1982. Judgment Under Uncertainty - Heuristics and Biases: A Progress Report on the Training of Probability Assessors. Cambridge University Press, Cambridge.

Arrow, K., 1986. Rationality of Self and Others. Rational Choice. University of Chicago Press, Chicago.

Arruñada, B., 2008. New Institutional Economics - A Guidebook: Part 1 Foundations: Human Nature and Institutional Analysis. Cambridge University Press, Cambridge.

Barberis, N., Thaler, R., 2003. Handbook of the Economics of Finance Volume 1 (Set): Chapter 18: A Survey of Behavioral Finance. Elsevier B.V., Amsterdam.

Berelson, B., 1952. Content analysis in communication research. The Free Press, Glencoe.

Broadbent, D., 1958. Perception and Communication. Pergamon Press, New York.

Camerer, C., 1995. Handbook of Experimental Economics: Individual Decision Making. Princeton University Press, Princeton.

Campbell, J., Lo, A., MacKinlay, A., 1996. The Econometrics of Financial Markets. Princeton University Press, Princeton.

DeBondt, W., Thaler, R., 1995. Handbooks in Operations Research and Management Science - Finance: Financial Decision-Making in Markets and Firms: A Behavioral Perspective. Elsevier, Amsterdam.

Edwards, W., 1968. Formal Representation of Human Judgment: Conservatism in Human Information Processing. Wiley, New York.

Fisher, I., 1928. Money Illusion. Adelphi, New York.

Gibson, G., 1889. The Stock Markets of London, Paris and New York. G.P. Putnam's Sons, New York.

Gilovich, T., Griffin, D., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, Cambridge.

Haugen, R., 1999. The New Finance: The Case Against Efficient Markets, 2nd Edition. Prentice Hall, New York.

Haugen, R., Lakonishok, J., 1988. The Incredible January Effect. Dow Jones-Irwin, Homewood.

Hawawini, G., Keim, D., 1995. On the predictability of common stock returns: Worldwide evidence. In: Jarrow, R. (Ed.), Handbooks in Operations Research & Management Science, Volume 9. Elsevier Science, Amsterdam, pp. 497-544.

Keynes, J., 1926. The End of Laissez-faire. Hogarth Press, London.

Keynes, J., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Kindleberger, C., 1978. Manias, Panics and Crashes, 5th Edition. Wiley, New Jersey.

Kuhn, T., 1970. The Structure of Scientific Revolutions, 2nd Edition. University of Chicago Press, Chicago.

Lo A., MacKinlay, A., 1999. A Non-RandWalk Down Wall Street. Princeton University Press, Princeton.

Lo, A., 1997. Market Efficiency: Stock Market Behavior in Theory and Practice, Volume I and II. The International Library of Critical Writings in Financial Economics, Edward Elgar Publishing, Cheltenham.

Lowenstein, R., 2002. When Genius Failed: The Rise and Fall of Long-Term Capital Management. Fourth Estate, London.

MacLean, P., 1990. The Triune Brain in Evolution: Role in Paleocerebral Function. New York Plenum Press, New York.

Miller, M., 1991. Financial Innovations and Market Volatility. Blackwell, Cambridge.

Ritter, Jay., 2003. Handbook of the Economics of Finance Volume 1 (Set): Chapter 5 Investment Banking and Securities Issuance, Edited by G M Constantinides, M Harris and R Stulz. Elsevier B.V., Amsterdam.

Savage, L., 1964. The Foundations of Statistics. Wiley, New York.

Schwert, G., 2003. Handbook of the Economics of Finance Volume 1 (Set): Chapter 15: Anomalies and Market Efficiency, Edited by G M Constantinides, M Harris and R Stulz. Elsevier B.V., Amsterdam.

Shiller, R., 1992. Market Volatility. MIT Press, Massachusetts.

Shiller, R., 2000a. Irrational Exuberance. Princeton University Press, Princeton.

Shleifer, A., 2000. Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press, New York.

Stone, P., Dexter, C., Marshall, S., Daniel, M., 1966. The General Inquirer: a Computer Approach to Content Analysis. MIT Press, Cambridge, MA.

Thaler, R., 2000. Choice, Values and Frames: Mental Accounting Matters. Cambridge University Press, Cambridge.

Tversky, A., 2004. Preference, Belief, and Similarity: Selected Writings. MIT Press, Massachusetts.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton.

Internet-based sources

General Inquirer (Harvard IV) word lists
(http://www.wjh.harvard.edu/~inquirer/homecat.htm)

Kenneth-French portfolios and their breakpoints
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Loughran and McDonald word list downloaded from
http://www.nd.edu/~mcdonald/Word_Lists.html

9 APPENDIX

Appendix A – History of the efficient market hypothesis

In this appendix, we aim to give the reader a quick review of the history of the EMH before the 1960s when the theory was introduced by Eugene Fama. A review of the history of the EMH might seem superfluous at first glance. However, in the words of John Maynard Keynes (1926):

> "A study of the history of opinion is a necessary preliminary to the emancipation of the mind"

The roots of the EMH date back far beyond 1965. The EMH is, and has always been, closely linked with the random walk -theory. In fact, the random walk -theory can be considered as an extension of the EMH (Fama, 1970).[203] Random walk asserts that successive price changes (or returns) are independent and identically distributed whereas the '*fair game*' models; also known as the expected return models, of the EMH state that prices (returns) '*fully reflect*' the information-set in question which implies that successive prices (returns) are independent of each other. Subsequently, the '*fair game*' models restrain from taking a stand concerning the form of successive distributions. In fact, the stochastic process generating returns is left unanswered in the EMH.

The history of the random walk -theory predates the EMH. Random walk -theory is based on the findings of a Scottish botanist Robert Brown who found that grains of pollen suspended in water had a rapid oscillatory motion when viewed under microscope (Brown, 1828). The finding of Robert Brown was later named as Brownian motion, the phenomenon used to model stochastic stock price movements among other things. In 1880, Thomas Thiele described the mathematical properties of Brownian motion (Thiele, 1880). Independent of Thiele's work, Lous Bachelier (1900) derived the mathematical and statistical properties of Brownian motion in the context of stock and option markets in his renowned PhD thesis: Theorie de la speculation. However, Bachelier's work attracted little academic attention at that time and it was until the 1960s when Bachelier was awarded the credit he deserved. Unaware of Bachelier's work, Albert Einstein (1905) developed the mathematical properties

[203] Indeed the early studies often involved testing random walk -models when in fact it can be shown that they were testing for a more general 'fair game' -model (Fama, 1970.

of Brownian motion and introduced it to the field of physics. From there on, Brownian motion, and the following random walk -models, have been widely used and studied.

Also the EMH assertion that prices '*fully reflect*' all available information dates back beyond Fama's introduction of the hypothesis in 1965. In 1889, George Gibson wrote the following: 'shares become publicly known in an open market, the value which they acquire may be regarded as the judgment of the best intelligence concerning them' in his book entitled: The Stock Markets of London, Paris and New York (Gibson, 1889). The similarity in Gibson's statement vis-à-vis the EMH is striking. Furthermore, the renowned economist John Maynard Keynes has been linked with the roots of the EMH. In 1923, Keynes argued that investors on financial markets are rewarded for bearing the risks associated with their investment instead of having privileged information compared to the market (Keynes, 1923). Keynes (1936) expressed a similar line of thought in his milestone work: The General Theory of Employment, Interest and Money. Besides Keynes and Gibson, Alfred Cowles III was a key figure in the history predating Fama's introduction of the EMH. In 1933, Cowles analyzed the performance of investment professionals and concluded that stock market forecasters cannot forecast (Cowles, 1933). In 1944, in continuation of his previous work, Cowles reported that investment professionals cannot beat the market (Cowles, 1944). In addition to the supportive evidence for the upcoming EMH, Cowles was the only author, to our knowledge, to publish evidence contradicting the EMH before the 1960s by pointing out inefficiencies in the market in the form of serial correlation in averaged time series indexes of stock prices (Cowles and Jones, 1937). However, Working (1960) showed that the use of averages can introduce autocorrelations not present in the original series.[204] In response to the critique, Cowles revisited his previous conclusions and corrected the error caused by averaging. However, Cowles still found mixed temporal dependence in results (Cowles, 1960) contradicting the forthcoming EMH and thus remains, to our knowledge, as the only author to publish contradicting results to the EMH before the 1960s.

[204] Independent of Working (1960), Alexander (1961) finds that averaging could introduce spurious autocorrelation.

Appendix B – Main variable definitions

This appendix provides definitions for the main variables used in the paper

Dependent Variables	**Description**	**Reference Literature (e.g.)**
Abnormal returns [0,1] [1,5] [2,32] [2,62]	Returns are based on buy-and-hold approach for the event period using close-to-close prices. Each stock is matched with 1 of 25 book-to-market and market equity portfolios at the end of June based on their respective market capitalization at the end of June and book equity of the last fiscal year-end in the prior calendar year divided by the market value of equity at the end of December of the prior year. The daily returns of the 25 size / Book-to-market portolios are retrieved from Kenneth French's website.[1]	Demers and Vega, 2010 Hirsleifer et al., 2009 Engelberg, 2008 Chan, 2003 Barber and Lyon, 1997 Daniel and Titman, 1997 Fama and French, 1992
Abnormal volume [1] [2,5] [3,32] [3,62]	Abnormal volume is defined as the log of the sum of the daily abnormal volumes for the event period. Daily abnormal volume is defined as: (day's volume - mean volume) / standard deviation of volume. The mean and standard deviation estimates are based on days [-65, -6] from the event date.	Loughran and McDonald, 2011 Hirsleifer et al., 2009 Tetlock, 2007 Antweiler and Frank, 2006
Abnormal volatility [2,32] [2,62]	Abnormal volatility is defined as the standard deviation of daily abnormal returns for the event period.	Loughran and McDonald, 2011 Demers and Vega, 2010 Engelberg, 2008
Independent Variables		
Sentiment LPS MPQA Wordcount Finance dictionary H4N dictionary	The quantitative estimate of investor sentiment. We estimate the value using alternative approaches LPS, MPQA and bag-of-words (wordcount.) We aggregate multiple daily news with equal weighting via averaging in both methods.	Loughran and McDonald, 2011 Demers and Vega, 2010 Davis et al., 2008 Engelberg, 2008 Tetlock et al., 2008 Tetlock, 2007
Market news volume	Market news volume counts the number of news in our sample for a given day to establish a proxy for distraction	Hirsleifer et al., 2009
Firm news volume	Firm news volume counts the number of news in our sample for a given day for a given firm	Loukusa, 2011 Fang and Peress, 2009

[1] Retrieved from Kenneth French's website on July 2012: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Appendix C - Main specification control variable definitions

This appendix provides definitions for the control variables used in the paper

Control Variables	Description	Reference Literature (e.g.)
Size	Log of market equity defined as the number of shares outstanding times the price of the share.	Loughran and McDonald, 2011 Hirsleifer et al., 2009 Tetlock et al., 2008
Book-to-market	Log of last reported book value of equity divided by the market value of equity.	Hirsleifer et al., 2009 Tetlock et al., 2008
Momentum [-4,-1] [-34,-4] [-255, -34]	We control for firm's past returns by utlizing a buy-and-hold approach with close-to-close prices to calculate abnormal returns for the given time intervals before the event date. Abnormal returns are calculated in the same manner as the dependent variable abnormal returns.	Loughran and McDonald, 2011 Tetlock et al., 2008 Tetlock, 2007 Chan, 2003 Jegadeesh and Titman, 1993
Share turnover	Log of the sum of the volumes for [-252, -2] divided by shares outstanding at event date.	Loughran and McDonald, 2011 Tetlock et al., 2008
SUE	We define Standardized Unexpected Earnings as the difference between the last reported quarter's EPS and the corresponding median analyst forecast for that EPS divided by the closing share price on the day of the respective earnings announcement.	Loughran and McDonald, 2011 Demers and Vega, 2010 Hirsleifer et al., 2009 Davis et al., 2008 Tetlock et al., 2008 Li, 2006
Abnormal volatility [-252, -2]	Abnormal volatility is defined as the standard deviation of daily abnormal returns for the time period.	Loughran and McDonald, 2011 Demers and Vega, 2010 Engelberg, 2008
Institutional ownership	The most recent number of shares reported under the ownership of institutions divided shares outstanding	Loughran and McDonald, 2011 Hirsleifer et al., 2009 Engelberg, 2008
Abnormal market volume [1] [2,5] [3,32] [3,62]	Abnormal volume is defined as the log of the sum of the daily abnormal volumes for the event period. Daily abnormal volume is defined as: (day's volume - mean volume) / standard deviation of volume. The mean and standard deviation estimates are based on days [-65, -6] from the event date. Daily market volume is based on the sum of the volume of the SP 100 firms.	Loughran and McDonald, 2011 Hirsleifer et al., 2009

Appendix D – Alternative specification control variable definitions

This appendix provides definitions for the additional control variables used in the paper

Control Variables	**Description**	**Reference Literature (e.g.)**
Industry	Fana and French (1997) 48 different industry dummies	Loughran and McDonald, 2011 Fama and French, 1997
# of analysts following	Log (1 + the number of analysts following a given firm)	Demers and Vega, 2010 Hirsleifer et al., 2009
Analyst dispersion	Standard deviation of EPS forecasts for the most recent reported EPS divided by the share price of the respective earnings announcement date	Loughran and McDonald, 2011 Demers and Vega, 2010 Tetlock et al., 2008
Calendar	Day of the week, end of the month, and January dummies	Hirsleifer et al., 2009 Tetlock, 2007
LTM Dividend	Last twelve month dividends divided by the last reported book value of equity	Fama and French, 2006 Li, 2006
No LTM dividend	No dividend dummy for firms that have paid no dividends during the last twelve months	Fama and French, 2006

Appendix E – Outlier values in financial data

We check the dataset for outliers. List of share price changes in the sample above +50% and below -50% are listed in the table below.

Date	Company	Return	Explanation
11/24/2008	CITIGROUP	58 %	Citigroup bailout package announced during the financial crisis
10/13/2008	MORGAN STANLEY	87 %	Morgan Stanley gets a significant investment during the financial crisis

Additionally, we list all exceptionally large jumps in market capitalization. Due to M&A and capital structure changes (changes in number of shares) these are not equivalent to the share price changes. Market capitalization changes in sample above +50% and below -50% are listed in the following table.

Date	Company	Change	Potential explanation[205]
4/8/2008	ALTRIA GROUP	-70 %	Kraft Foods to spinoff (actually split-off) Post Cereal that will merge with a Ralston company
6/12/2006	LOWE'S COMPANIES	-50 %	Potential change in capital structure related to stock split (jump does not exist for stock price data)
12/29/2006	AT&T	63 %	Completion of acquisition of BellSouth Corporation
7/2/2007	BANK OF NEW YORK MELLON	59 %	Merger of BNY and Mellon on Jul 1; jump has been corrected in stock price data of Datastream
11/26/2008	CITIGROUP	58 %	No clear event, but change visible also in stock price. Potential correction (Citi stock price down from >$130 to <$40 in less than a month
7/30/2009	CITIGROUP	101 %	Citi Announces Final Results of Public Share Exchange and Completes Further Matching Exchange with U.S. Government; Citi Pushing for Quick Asset Sales
9/14/2009	CITIGROUP	96 %	Potentially a change in capital structure (jump has been corrected in stock price data of Datastream)
3/22/2007	CVS CAREMARK	84 %	Merger of CVS Corporation and Caremark Rx Inc.
3/19/2007	FREEPORT-MCMOR.CPR.& GD.	66 %	Freeport-McMoRan Copper & Gold Inc. acquired Phelps Dodge Corporation, including Chino.
4/2/2007	KRAFT FOODS	246 %	Change related to Kraft spinoff by Altria during Mar 2007
3/7/2006	LOWE'S COMPANIES	96 %	Potential change in capital structure related to stock split (jump does not exist for stock price data)
1/1/2007	MASTERCARD	69 %	Potentially a change in capital structure (jump has been corrected in stock price data of Datastream)
10/13/2008	MORGAN STANLEY	87 %	Mitsubishi UFJ Financial Group Closes $9 Billion Equity Investment in Morgan Stanley as Part of Global Strategic Alliance
3/14/2005	NATIONAL OILWELL VARCO	65 %	National Oilwell Varco emerges (from Oilwell Supply,1862 and National Supply, 1894) following the completion of the merger with Varco International
8/15/2005	SPRINT NEXTEL	102 %	The Sprint/Nextel merger was finalized
3/31/2008	VISA 'A'	97 %	Potentially an error in the data (jump one week after listing)
2/5/2010	XEROX	53 %	ACS Signs $1.6 Billion California Medicaid Contract

[205] Source: internet search

Appendix F – Details on gathered media data

This appendix explains in greater details how we have gathered our text data from the LexisNexis database. Due to the technical nature of the process, we have excluded this from the main section. The appendix details the list of excluded news sources, the looks of the database interface, and the process of downloading data.

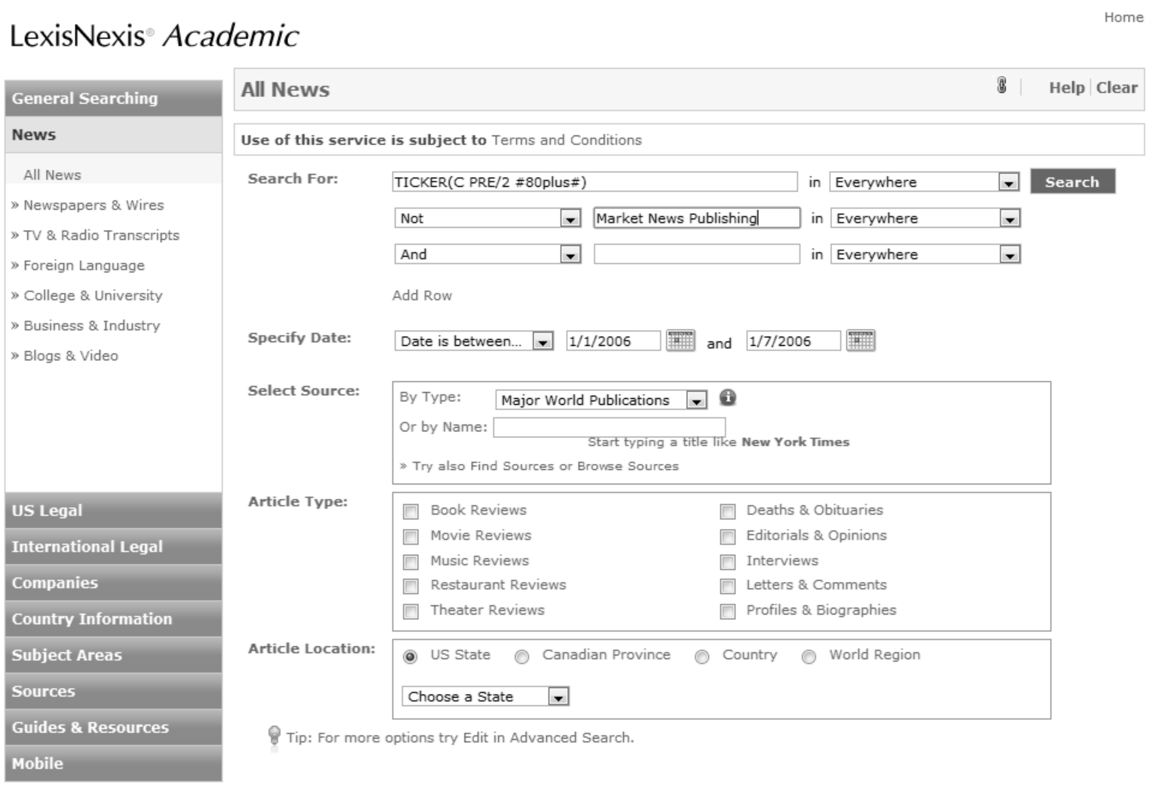
List of excluded sources

The table below lists the excluded data sources and rationale for excluding each.

Excluded source	Rationale for exclusion
Market News Publishing	Only numerical data
GlobalAdSource (English)	Provides information on company advertisements in a tabular form
News Bites - Nordic : Finland	Provides automatically generated updates of stock price developments etc.

LexisNexis search interface

The following screenshot illustrates the LexisNexis user interface that we use to download the news with the web scraper. As can be seen in the example, we need to split our search for large companies to multiple small pieces to stay below the 500 news items limit.



Use of this service is subject to Terms and Conditions

Search For: TICKER(C PRE/2 #80plus#) in Everywhere
Not | Market News Publishing in Everywhere
And | in Everywhere

Specify Date: Date is between... 1/1/2006 and 1/7/2006

Select Source: By Type: Major World Publications
Or by Name: Start typing a title like **New York Times**
» Try also Find Sources or Browse Sources

Article Type:
□ Book Reviews □ Deaths & Obituaries
□ Movie Reviews □ Editorials & Opinions
□ Music Reviews □ Interviews
□ Restaurant Reviews □ Letters & Comments
□ Theater Reviews □ Profiles & Biographies

Article Location: ● US State ○ Canadian Province ○ Country ○ World Region
Choose a State

Tip: For more options try Edit in Advanced Search.

Details of the web scraper

We download our news from the LexisNexis database. As it is not possible to download more than 500 news at a time from the database, we develop a web scraper that imitates the actions of a human user navigating the LexisNexis website and downloading the news in small portions at a time. We develop our program using *AutoHotkey*, a free open-source macro program that can be used to automate different keyboard and mouse usage on a Windows computer. We program the scraper to imitate our mouse and keyboard movements: entering a date range and a ticker, and then downloading the news items.

When searching for news, the number of search results depends on the time period. If the number of search results exceeds 500, LexisNexis lets us download only the first 500 news items, and we would miss the rest. Thus, we need to keep the number of found news items always below 500. In order to achieve the aforementioned, we need to split the whole time period into smaller pieces for each company, for which we are certain that the amount of news found never exceeds 500. For a company with less than 500 news during a period of 5 years, we can download all the news at once, while for companies with more news coverage, the frequencies need to be very small. Downloading news with small frequency for all the companies would not be practical, as downloading time is always constant. For example, using a 3-day time period would need 640 downloads per company[206]. Given that our web scraper can complete one download in approximately three minutes, a total download time would equal to 640*3 minutes = 32h per company, or over 133 days for the whole sample. Thus, we need to determine the largest possible time periods per company for which we are confident that the number of search results will not exceed 500 items.

The amount of news varies considerably among the sample: the smallest companies have lower news coverage; however, the larger companies also have significant variation in their news volume. Thus, we do a sample search for a period of one year per company, and based on these results, we estimate as a rule of thumb that the maximum volume that can occur in a certain timeframe is three times the frequency. For example, with a test search of 400 news items in a year, our estimate of the number of maximum news per that company would be 1,200 per year, and hence would dictate a time frequency of 4 months to not exceed the 500

[206] As the speed of a webpage loading varies significantly depending on the number of found news items, the program needs to be set to wait for a webpage to load for a longer time than a human user would wait for. Otherwise, the scraper would start navigating on a page that has not been loaded yet. This causes the scraper to be slower than a human user would be. Furthermore, the download time does not correspond to the amount of news downloaded, but only to time period.

limit per download. However, instead of using arbitrary time period based on the 1 year sample news volume, we define constant time periods that we apply based on the 1 year sample period. Therefore, in our example case, we would use time period of 3 months to not exceed 500 news per year. The time periods we use are: all news at once (e.g. Weyerhauser Co.), one year per download, 6 months by download, 3 months by download, 1 month by download, 2 weeks by download, 5 days per download, and 3 days per download (e.g. Citi, Google and Apple). By doing this, we are able to cut our total download time significantly. To further speed up the downloading process, we edit our scraper to split our download queue so that we can use multiple computers.

Besides being fairly inefficient, another peril of a human-simulating scraper is the possibility of distractions. Other open programs that provide pop-ups, or unexpected search results on a webpage, may result in unexpected consequences, as the computer program keeps clicking on the wrong window. In the worst case scenario, this can lead to a scraper navigating itself to wrong places on a website - potentially even posing a threat to the whole computer. Therefore, it is not possible to leave the scraper fully unsupervised for long periods of time.

We finally launch our scraper with total of 50 computers that we supervise as they download the whole sample within a few days. Once we have downloaded data for all the companies, we study the download results to find companies that may have exceeded 500 items. In these rare cases, we download the missing news manually. As an outcome, we have a folder for each of the SP100 companies, containing a number of HTML-files that represent less than 500 news items per search.

Details of processing LexisNexis data

After the web scraper process, we have a sample of 5,389 HTML-files that contain all the news for our time period that have been marked with appropriate tickers in LexisNexis database. As all metadata: e.g. dates of news, publications, etc., are stored in these HTML-files, we cannot proceed further without processing of the files more.

To do so, we design a Java-algorithm that puts all news items in the files into a database in a common format. The program starts by taking a HTML-file and cutting it into individual stories. Next, the algorithm needs to discover the heading, the text, the publication, and other metadata. As LexisNexis HTML format is non-standard, it does not provide a clear indication of which part is what. Therefore, we program the algorithm to interpret the possible meaning of a text from its relative position to others: e.g., the date comes before the heading etc. Sometimes there is other metadata between the classified items that does not match the standard ordering: i.e., date follower by heading etc. In such cases, the algorithm verifies the classifications by looking at the forms of the data: i.e., the words that start a sentence, the sentence structure and other format related things.

Once the message and other metadata have been separated, we strip the HTML-tags off from the message to make it easier to process in the data analysis phase. While doing this, we also remove tables as they contain mainly numerical data that is not interesting for us. The exceptions are messages that are fully expressed in a table format: in this case, we only remove the html-tags and leave the content of the tables. We also alter some of the metadata, in particular the date field. The date format varies occasionally, and our algorithm recognized most of the date formats. Some messages are; however, discarded, as the date is not properly specified. Such messages are mainly composed of situations where the date is in a format such as: ‘November 2009’. Such formats are not useful as we wish to calculate the sentiment for a particular date, and we cannot place a message to a particular date using the aforementioned format as it does not specify a specific date, but only a month and a year.

Once the structure has been discovered, we store the messages into the database including: dates, tickers, source publications, and other relevant data that LexisNexis provides. As the amount of data that we have at this point is very large, we choose to store the ordered news into a MySQL database that we setup on a home server. MySQL is a widely used relational database that is chosen due to its ability to handle large amounts of data, and its availability as a freeware. The stored data is then used as a basis for our analysis.

Appendix G – Annotation instructions

The goal is to indicate if the following news would likely affect a company's share price positively or negatively. If you think that the share price could go in either direction, please select 'Either way'.

Please consider all sentences in isolation, i.e., how do you expect the company's stock price to react if this was the only headline that you saw, and you had no background knowledge on the company, or current economy. I.e. we don't want you to make your own stock analysis. Expect that you know that the economy is currently steady and that the company is a stock-exchange listed company (i.e. large company).

When evaluating the sentiment of sentences and how the stock price would react, please use the following scale:

9	+++	*UP a lot*
8	++	*UP (hard to say how much)*
7	+	*UP a little*
6	n1	*Either way (answering would require more background knowledge)*
5	n2	*Either way (not possible to say, even if I had more background knowledge)*
4	n3	*Either way (answering would require me to know which company we are talking about)*
3	-	*DOWN a little*
2	--	*DOWN (hard to say how much)*
1	---	*DOWN a lot*

Appendix H – Financial entities -wordlists

In this appendix, we list show our financial entity wordlists, originated from the online dictionary Investopedia (for further details on retrieving this list, refer to section). The first list shows words where an increase would be typically considered a negative event. The second list covers the opposite, i.e. words where an increase would typically considered a positive event.

Negative-if-up word list

ABI
acquisition cost
add-on
administrative expenses
antitrust
bankruptcy
bear
beta
capital employed
capital loss
current assets
current liabilities
debt
default
deficit
dilution
downgrade
downside
downtrend
EAC
erosion
expense
financial crisis
gearing
gearing ratio
impairment
income tax
inflation
interest rate
inventory
leverage
liability
loan
long-term debt
net debt
net loss
noise
NWC
obligation
OPEX
overhead
profit warning
provision
receivables
recession
relative strength
restructuring charge
risk
RSI
seasonality
seller
sell-off
shortage
shortfall
slump
tariff
tax expense
tax rate
taxation
taxes
underperform
unemployment
volatility
working capital
write-down

Positive-if-up wordlist

AAA
ADR
asp
basic earnings per share
benchmark
bonus
book value
boom
bottom line
brand
budget
business
buyback
CAC
capacity
CAPEX
capital
capital gain
capital markets
capitalization
carried interest
cash
cash flow from operating activities
catalyst
closing price
competitive advantage
complement
composite index
comprehensive income
consumer spending
credit
credit rating
cash equivalents
cash flow
cash flow from investing activities
customer
customer service
DAX
delivery
demand
dividend
EBITDA
economic growth
economic recovery
economics
economy
efficiency
end-user
EPS
equity
exemption
export
financial market
financial
performance
financing
flotation
franchise
FTSE
Fundamentals
HAM
holdings
income
index
intangible asset
intellectual property
investment
dividend policy
earnings
EBIT
EBITA
IPO
leadership
liquidity
LSE
MACD
margin
market
market capitalization
market price
market sentiment
market share
market value
maturity
monopoly
net cash
net interest income
net investment
net sales
Nikkei
NYSE
offering
operating earnings
operating income
operating margin
operating profit
operational efficiency
order
outperform
partnership
payout
pipeline
plum
GDP
gross margin
gross profit
growth rates
population
portfolio
pre-market
price target
private placement
productivity
profit
profit margin
quotation
rally
rating
rebound
refinance
remuneration
repayment
retail sales
retained earnings
revenue
ROCE
ROE
ROI
rollout
royalty
savings
sequential growth
share capital
share repurchase
shareholder value
solvency
stock market
stock option
STOXX
subscription price
subscription right
supply
surplus
synergy
takeover bid
throughput
top line
trademark
TSE
TSX
turnaround
turnover
underlying asset
unit sales
upside
uptrend
valuation
value
volume
WAL
wall street
wealth

Appendix I – Wordlist defects

To test for possible wordlist defects, we go through the sentences that have been falsely assigned by our algorithm (vs. annotation). The most commonly occurring words from the word lists are listed below.

	Wrongly assigned	**Correctly assigned**
deal	71	22
acquire	55	12
support	41	12
agreed	35	7
completed	34	7
provides	32	6
enables	19	1
natural	19	3
approval	14	1

Appendix J - Error descriptions for LPS

After labeling our annotation sample with the Linearized Phrase-Structure -model algorithm, we categorize the cases in which the algorithm gives us false answers (different to human annotation). The definitions of these errors are as follows.

Borderline cases: sentences that could be tagged by human as either/or

Let us offer an illustration of such a case:

'*The long-standing partnership and commitment enable both parties to develop their respective operations, and ESL Shipping will also have the opportunity to update its fleet and improve its efficiency.*'

The sentence implies that the company's efficiency will increase. A human reading the sentence could think that the increase could be substantial enough for the share price to increase. Another reader could think that the aforementioned is mainly glitter and the increase is not substantial. Thus, two readers might annotate the sentence differently.

Company talking in advertising like -tone about its' own operations

In some instances, a company can talk in a very positive tone about its operations, and our algorithm picks this up; tagging the sentence as positive. An analyst, on the other hand, would often disregard this information, or at least discount it significantly. Therefore, an analyst is able to recognize a difference between objective product reviews and company's own review, for instance. A positive review can be good news for the company, while a company talking about its own operations in a positive way should often not be news to anyone.

Inability to detect changes in numbers

The algorithm is unable to detect that two different numbers in a sentence may mean an increase/decrease. For instance, for sentence:

'*The company recorded sales of 100 million, vs. 10 million in last year*'

the algorithm would not recognize that sales have increased significantly. A human annotator, on the other hand, could record this as a positive event.

Inability to detect roles in a sentence

The algorithm cannot detect roles of different entities in a sentence that may become problematic in cases when multiple parties are described in the sentence. For example, a sentence may talk about a company getting money from another company after a legal case: good news for one of the companies and bad for the other. As our algorithm does not know the point of view, it is prone to misclassify such sentences; hence, increasing noise in the sentiment.[207]

Inability to detect time expressions in a sentence

An event that has already happened may be described in a very polarized tone. While a reader can be impacted by simply the tone of the text, it is also possible that an investor would not give much weight to events of the past - regardless of tone of the text. The reaction depends on whether or not investors are impacted by tone, or by information, or by both.

Inability to reason from text

In cases where a reader makes a connection that is not explicit in the text, the algorithm is incapable of uncovering such reasoning through logic. For example, a construction company building something over the next years could imply that it has received a large order even though such statement is not explicitly expressed in the text.

Inability to recognize significance of events

In some instances, events are reported that are not likely significant enough to make a difference to a company. These can be tagged as polarized, though they are not significant enough to impact share price. The sentiment is correct, but the impact on the company is likely too insignificant. In such cases, an ideal algorithm should either tag the sentences as neutral, or at least adjust the score with a smaller weight during the aggregation. In some instances, news may include polarized sentences that have nothing to do with the company that the article covers.

[207] This category has similar elements to the neutrally annotated category ‘Either way’ (answering would require me to know which company we are talking about).

Inability to understand the magnitude or value of items

Our algorithm cannot understand magnitude. For example, if a ship yard gets ‘*a new order of 100 tankers*’, the algorithm would recognize no difference to an order of 100 candy bars. It would be evident to a reader that the candy bars are unlikely to present much value to a company. However, the algorithm is unable to separate between the value of a tanker and a candy bar; hence, both orders seem to be of equal importance.

Interpreting non-words as words

In some instances, our algorithm interprets an entity, or a non-word, as a word. For example, when talking about the firm ‘*Capital One Financial*’, the algorithm recognizes that ‘*Capital*’ is the word ‘*capital*’.

Need for more context

In many instances, a reader would need more context than one sentence in order to make any conclusions. For example, a sentence ‘*The CMO is thrilled that the acquisition is going through.*’ may sound positive. However, acquisitions may often destroy value, and an analyst would want to know the background of the acquisition before making any conclusions based on the information.

Polysemy of words and expressions

Sometimes we recognize words as they would be used in a different context. For example, when a company ‘*succeeds*’, this is good news. On the other hand, when a person ‘*succeed*s another the meaning is that a person is replaced by another; not positive or negative news in itself. In order to correctly tag these sentences, we should move from word polarity to ‘*sense polarity*’: i.e. recognizing that words may have different polarity in different context (e.g., Maks and Vossen, 2010).

Positive convention of talking about something

With some themes, there exists a potential bias of positivity that relates to a convention with a certain theme rather than the actual sentiment. In such instances, the algorithm can classify sentences as positive whereas in reality they are simply reflecting the convention of the theme. For example, nominations are often described in an overly positive manner as a matter of convention instead of actually reflecting facts.

Recognizing patterns in descriptive text

Often financial texts simply describe what operations of companies. In some instances, these descriptions may include words that are included in our wordlists. The algorithm may recognize a polarized pattern in such descriptive texts, even though they are in fact neutral by nature.

Sentences with multiple parts

The SVM is typically reading a sentence as a whole, ignoring punctuation that splits two parts. While the aforementioned does not generally matter, there may be instances where a sentence should be handled in two parts with different sentiment structures.

Use of longer expressions and interpreting non-words as words

Our algorithm can detect expressions that consist of only individual words. However, in some cases, multiple words form an expression that has a different polarity than individual words. For example, we may recognize '*limited*' as a negative word, while the text may describe actually a '*limited company*'. To mitigate this, there would need to be a list of common expressions and their polarities.

Words lacking from word lists

The algorithm recognizes only words that are included in our word lists. While our lists are long, and include the most commonly used polarized words, they may lack some words. As the words are not recognized, the SVM remains unable to pick up the polarized pattern.

Wrong computer patterns

The computer patterns that have been created from the training set may be erroneous: for example, movement words can be misclassified to represent sentiment when in fact they do not. This may be a result of lack of training data for some of the patterns. Alternatively, we may be lacking some word types from our word lists that result in misclassified annotations, and hence may lead into a bias in the training set. In example, if we did not have '*negate*' category in our training set, the set would be interpreted very differently by the SVM as positive words would lead to negative sentences etc.

Wrong label in training data

When doing our error analysis, we also notice instances where we have a wrong label in our training set. In other words, while our algorithm is able to classify the sentence correctly, the human annotator has made a mistake when tagging the sentence.