ABSTRACT

Title of dissertation: ESSAYS ON
NEWS AND ASSET PRICES

Nitish Ranjan Sinha, Doctor of Philosophy, 2010

Dissertation directed by: Professor Albert S. Kyle
Department of Finance
Robert H. Smith School of Business

The first essay examines news and the cross section of returns. Using a sentiment score provided by Thomson Reuters to measure the tone of news articles, this paper examines monthly portfolio returns constructed from information about past news articles. The sentiment score is obtained from the kind of words and phrases that are used in the news article. Positive tone in news articles in the past months predicts positive returns. Similarly, negative tone in the past months predicts negative returns. Past sentiment predicts future returns even for large stocks. The predictive ability of past sentiment dominates the predictive ability of past returns. After controlling for past sentiment, the predictive ability of past returns (in predicting future return) disappears. The findings are robust to multiple specifications. The predictive ability of past sentiment can be used profitably. When applied to the largest decile of stocks, a strategy that takes a long position in stocks with past positive sentiment score and a short position in stocks with past negative sentiment score generates a statistically significant alpha of 34 basis points per month.

The resulting portfolio is also positively correlated with a long-short momentum portfolio. Within the same time period, a trading strategy using the sentiment scores from the subset of news articles citing analysts is not profitable. The news items that cite analysts have economically significant contemporaneous returns. The findings suggest that (i) the market underreacts to information contained in news articles, (ii) momentum might be related to underreaction to the sentiment information, and (iii) market participants pay attention to sentiment score information in analyst news. The findings are consistent with a model where one trader has private information and others are trading based on past returns and volume information. The paper also shows that after adjusting for firm size, stocks with abnormally high counts of news articles underperform stocks with normal counts of news. Stocks with abnormally low newscounts also underperform.

The second essay examines the relationship between news and trading activity. The theory of trading game invariance of Kyle and Obizhaeva (2009) predicts that for every one percent increase in trading activity, the frequency of news articles should increase two-thirds of one percent. Using news data from 2003 to 2008, we show that the cross-sectional variation in news articles across stocks is related to the trading activity in a manner consistent with the trading game invariance. The relationship is robust to various estimation procedures including models of count data. The relationship is also robust to multiple ways of counting news and excluding various type of firm specific news.

ESSAYS ON NEWS AND ASSET PRICES

by

Nitish Ranjan Sinha

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Albert S. Kyle, Chair/Advisor
Professor Russ Wermers
Professor Gerard Hoberg
Professor Michael Faulkender
Professor Philip Resnik

Acknowledgments

One can be lucky enough to have an advisor who gives good advice. I am extremely lucky since I got the same good advice from all my trusted advisors - my parents, my brother Nihar Ranjan, my wife Elisabeth Newcomb Sinha and my thesis advisor Pete Kyle. I am thankful to all of them. They all encouraged me to do what I would find satisfying. My parents watched me grow and were often disappointed by my propensity to take on projects bigger than my size. They learned to offer the same warning before I started a project -"Make sure you can finish it." Elisabeth and Pete offered the same warning when I started out on this project. Elisabeth also did much more than offering advice, but I have a lifetime to thank her for all that, so I will save space in this section by leaving most of it out. Talk about being lucky.

I am thankful to my advisor Pete Kyle for encouraging me to pursue this line of research. Early in the program, Pete spent countless hours sifting through my ideas. Later he spent countless hours sifting through drafts and tables for discussion. Despite his really busy schedule, Pete was always available. He also encouraged me to be an independent researcher and not be afraid of my limitations but think about the possibilities. Thank you Pete. I am also grateful to his wife, Katherine Kyle for sharing Pete's time so generously with me.

I am also thankful to Jerry Hoberg, Russ Wermers, Mike Faulkender and Phil Resnik for their guidance as committee members. Many thanks to Rich Brown and William Fang at Thomson Reuters and Chuck LaHaie at the University of Maryland

for ensuring access to the data. Rich also spent countless hours explaining the nuances of the news data. I am grateful to Fabrice Courbon, product manager Machine Readable News at Thomson Reuters for his patience with me. He was always available for a clarification on news data. I am also grateful to Giles Mayley at Infonic for going over the text processing algorithm despite his tight deadlines. The dissertation benefited from discussions with Steve Heston, Dalida Kadyrzhanova, Anna Obizhaeva and Tugkan Tuzun. Anna's, Dalida's and Steve's comments helped me with the presentation of the first paper. The second essay would not have been possible without help from Anna and Tugkan.

My research has also benefited from the general research climate at Maryland, particularly the summer presentations organized by Gordon Phillips. Gordon also set me up for making a living as an empirical finance guy through his 828 course. Thanks Gordon! I am grateful to Nagapurnanad Prabhala and Georgios Skoulakis, though our discussions regarding volatility and news did not end up making into the dissertation. I felt enlightened and encouraged by their discussions. I was also encouraged by my discussions with John Rust in the economics department. John also introduced me to SQL databases in his computational econometrics course. My fellow graduate students, Jeongmin Lee, Wei Li, and Katie Moon helped me polish the writing by poring over the preliminary draft of my job market paper. I cannot be thankful enough to them.

I also benefited from discussion with S. Viswanathan at Duke University and Paul Tetlock at Columbia University. Paul Tetlock's and Vish's papers also set the stage for this kind of work for years to come. Thanks Paul, Thanks Vish!

Thanks to Mark Lowenstein for asking "How are things?" and expecting an answer every time. In my last five years at Maryland, I have been lucky to have Aysun Alp, Minwen Li, Su Li, Matt Kozora, Nitin Kumar, Anshuman Sinha and Yue Xiao as my classmates. They improved my emotional well being by being such good friends. I will miss you. During my five years at Maryland, I was beneficiary of kindness of David Hunter, Yun Liu, and Michael Padhi. Thanks to them too.

Madhur Singhai and Mandy Du from NYU helped me with setting up a parallel database of news articles at Thomson Reuters. I learned a lot from Madhur about setting up large databases. Mandy provided me with more opportunities to learn about the data. Reuters sales force exposed me to their clients allowing me to have an insight into how fund managers think about news. Thanks to them too!

Research is full of serendipity and dead ends. I met a dead end in form of my second year paper on mergers. I was particularly concerned by the problem of rare events while studying mergers. It was a comment from Marc Nerlove - "Rare events are often very informative, when they occur" - that got me thinking about studying the microstruture issues surrounding news items. Serendipitously, at the same time I was thinking of studying news items with Doron Avramov. Thanks to Marc, this dissertation is a step towards studying microstructure surrounding news articles, though it is not really studied in these two essays. Thanks to Doron for getting me started in this direction. Marc also demystified econometrics through his multiple AREC 699s. Every fall, we would meet on Mondays and Marc will start a discourse on a topic in econometrics. It typically consisted of problems that scientists were dealing with at the time of this particular breakthrough and the dominant paradigm

at that time. In other words, he would set each topic within its context of discovery. That helped a lot! Pretty soon things fell into a pattern and things started to make sense. I can only wish every PhD student to have access to such seminars.

Working on the dissertation has been an exhilarating experience. It was made extremely memorable by absence of the "M" key on my computer while working on the essay on momentum. Thanks to my son Rohan Sinha, for taking the "m" out of momentum! Thanks to my in-laws and my parents for asking often enough about being done.

I have also benefited from the prevalence of research-quality code generated by the open source movement especially MySQL, Python and Unxutils. Same goes with fragments of SAS and STATA code all over the Internet and mailing lists. The dissertation is formatted using LaTeX. The template for the thesis came from Dorothea F. Brosius at the Institute for Research in Electronics and Applied Physics. Thanks to all the people who made their work public, thereby making my life easy.

I am grateful to my Uncle Steve Newcomb for discussing coding in Python with me. Over the course of my interaction with him, my python code improved. About now, I am ready to take on bigger, hoarier challenge in text processing. Thank you Aunt Vicky Newcomb for emacs and wget tips.

I also benefited from discussions in University of Alabama Tuscaloosa, University of Illinois Chicago, University of South Carolina and University of Washington Seattle. Thank you all seminar participants. Thanks to the seminar participants at Goldman Sachs Asset Management and Soros Fund Management.

I am also thankful to those whose kindness has made my life easy and whose

names are omitted by my sloppy memory. Finally I think I have been blessed in many ways. Thanks for all the blessings.

CONTENTS

LIST OF FIGURES

1. NEWS ARTICLES AND MOMENTUM

1.1 Introduction

Is news incorporated into prices quickly? Is slow incorporation of information related to momentum? What explains the speed of incorporation of news into asset prices? Is the news merely reflective of pessimism in the market or it is informative? This paper addresses these questions using a database of all news items that appear on a trader's screen through Thomson Reuters.

News events appear to affect stock prices quickly, suggesting that the market incorporates information fast. Very few studies have, however, undertaken a comprehensive analysis of how news is incorporated into prices.[1] Only recently has it become practical to quantify the content of a large set of news articles, allowing for the measurement of the reaction to positive and negative news over a long time period. This paper exploits a Thomson Reuters' firm-level measure of news articles' tone , or sentiment score. The sentiment score is derived from the words and phrases used in the article to describe the firm. If there are multiple firms

[1] Berry and Howe (1994), Mitchell and Mulherin (1994) and, more recently, Tetlock, Saar-Tsechansky, and Macskassy (2008) are important papers that have undertaken comprehensive studies of how news is incorporated in stock prices. While Berry and Howe (1994), and Mitchell and Mulherin (1994) study the arrival of news and its impact on volume, Tetlock, Saar-Tsechansky, and Macskassy (2008) focuses on how the textual information is incorporated into prices.

mentioned in an article, the score is specific to each firm. The analysis shows that past news sentiment score has predictive ability for future returns in addition to the predictive ability of past returns. The predictive ability of past returns (to predict future returns) disappears once past sentiment scores are incorporated in return expectation. The predictive ability of news sentiment score is economically significant. We examine a trading strategy based on sentiment score taking a long position in stocks with past positive sentiment score and a short position in stocks with past negative sentiment score. The portfolio is held for five months.[2] From 2003-2008, a portfolio of large stocks exploiting the trading strategy generates an alpha of 34 basis points per month after controlling for the three Fama-French factors and momentum factor. The horizon over which the trading strategy is profitable suggests that the market incorporates news information into stock prices rather slowly.

In addition, the sentiment portfolio is correlated with a momentum portfolio, suggesting that the slow absorption of news sentiment may explain some portion of momentum.[3] Jegadeesh and Titman (1993) suggest that momentum could be due to underreaction to information. In the finance and accounting literature, while there are other phenomena (e.g. post earnings announcement drift (Ball and Brown 1968) and the accrual anomaly (Sloan 1996)) that demonstrate that financial markets absorb information slowly, this is one of the first studies that show the relationship between slow incorporation of tangible information and the momentum phenomenon.

[2] The results are qualitatively similar for 1,2,3 and 4 months.

[3] Jegadeesh and Titman (1993) document momentum, the profitable trading strategy of selling stocks with poor return in the past 3-12 months and buying stocks with good return during the same time period.

Finance literature shows that two aspects of news are informative - the amount of news and the tone of news. This paper focuses on the tone of news. This paper continues the line of research pursued by Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Engelberg (2009) and Demers and Vega (2008), where the diction of news articles or press releases is analyzed. There are, however, some key differences. First, the unit of analysis is a sentence rather than words. Sentence as unit of analysis renders the sentiment score valid at the firm-news pair. If many firms are mentioned in the same article, the sentiment score is from the sentences specific to each firm. Second, the profitability of the sentiment-score based trading strategy indicates that the effect of the tone of news articles lasts longer than reported by Tetlock (2007), where the tone of Wall Street Journal's daily column "Abreast of the Market" is shown to affect the market next day. Tetlock (2007) analyzes a market commentary column, while this paper analyzes firm specific news articles. Similarly Tetlock, Saar-Tsechansky, and Macskassy (2008) find an abnormal return of 11.3 basis points at the daily level rather than abnormal returns persisting over months. Third, unlike Tetlock, Saar-Tsechansky, and Macskassy (2008) who report that the returns from news-based trading are not strongly related to any of the Fama-French factors or the momentum factor, the sentiment score portfolio correlates with the UMD factor. [4] Among the largest 600 stocks, the correlation is almost 62 percent.

Chan (2003), Fang and Peress (2008) and Akbas, Kocatulum, and Sorescu (2008) have studied the effect of the amount of news on stock returns. However

[4] UMD is the return of a portfolio of stocks with high past twelve month return, minus return on a portfolio of stocks with low past twelve month return.

they create long-short portfolios of stocks with news and no news. Conditioning on news versus no news is likely to result in very small number of large stocks in the sample. As reported here, large firms almost always get some news; hence, it is reasonable to make the distinction between usual and unusual levels of news.[5] To compare the amount of news across size groups, a modified measure of the amount of news - size adjusted newscount - is considered. Stocks with abnormally high or low size-adjusted newscounts underperform stocks with normal newscounts. A strategy which is long the 4,5 and 6 and short the 1, 2, 9 and 10 decile based on size-adjusted newscount generates an excess return of 45 basis points per month after adjusting for three Fama-French factors and momentum.

The two attributes of news items — sentiment score and size-adjusted newscount — have different relationships with momentum. The portfolio based on the sentiment score is positively correlated with UMD. In contrast, size-adjusted newscount has a non-monotonic relationship with momentum. The monotonic relationship between sentiment portfolio returns and UMD suggests that while the market might be underreacting to various kinds of information, an important component of the momentum effect is related to underreaction to the sentiment information in news articles.

Qualitative information is hard to process and hence is not evident to all market participants. Engelberg (2009) points out that the difficulty in processing qualitative information could be a source of this delayed reaction to the qualita-

[5] In the sample, firms which do not get any news in a month have an average size of 277 million USD. By comparison, 277 million is smaller than the average firm. The average of size decile 5 is 290 million USD.

tive information. However, this paper – along with Tetlock, Saar-Tsechansky, and Macskassy (2008), Engelberg (2009) and Demers and Vega (2008) – shows that it is profitable to parse qualitative information. It is likely that some investors already specialize in processing it. This paper examines predictions from various models of information incorporation as well.

1.1.1 Models for incorporation of information

Prediction		**Strategic trading based model**	**Overconfidence based model**	**Bounded rationality based model**	**Rational inattention based models**
Autocorrelation in returns followed by public event	Small stocks	Yes	No	Yes	Yes
	Large Stocks	Yes	No	Yes	Yes
Return predictability from previous returns dominated by qualitative information	Small stocks	Yes	Yes	Yes	Yes
	Large stocks	Yes	No	No	No
Autocorrelation in returns followed by public event that is conditional on misvaluation		No	Yes	Yes	Yes

Fig. 1.1: Predictions of various theories of incorporation of information

Strategic trading, limited processing ability of agents, and market frictions are possible explanations for why asset prices incorporate information slowly. Agents can have limited processing ability either due to rational inattention or behavioral biases. The evidence presented in the paper weighs in favor of strategic-trading-based explanation. Predictions from various model are summarized in figure 1.1.

Strategic trading based explanation

Foster and Viswanathan (1994) model trading when the information is long lived but is known to only one informed trader. Other traders infer the information from the order flow and returns in a Kyle (1985) framework. When the news arrives, the common information is quickly incorporated into price but the information known only to the informed trader (e.g. Warren Buffett) is slowly incorporated into prices. One implication of the model is that conditional on the information set of uninformed (but rational) traders, the returns are not predictable. However, the model of Foster and Viswanathan (1994) would suggest that conditional on the information set of the informed trader, returns are predictable. The profitable nature of a sentiment-based trading strategy even among larger stocks suggests that the sentiment information is closer to the information set of the informed traders. Additionally, in the Foster and Viswanathan (1994) model, the information set of an informed trader is superior to the information set of the uninformed trader. I find that the predictive ability of past returns is dominated by the sentiment information.

One does wonder whether the information in public news can be considered as private information. The implication of a public news item is almost always open to interpretation. Savvy investors are known to parse public news more carefully than the average investor. Let us consider the example of Warren Buffett, who reads 10-Ks before making investments. Famously, reading Lehman Brothers' 10-K helped Buffett decide not to buy at what appeared to be a bargain price. According to the Lehman Brother Holdings bankruptcy report, he turned down the deal on March

28, 2008. [6] Lehman Brothers filed for bankruptcy on September 15, 2008. Buffett was able to use his expertise in processing qualitative information to infer something six months before the market did. He also kept the interpretation to himself.

Behavioral biases based explanation

Investors could be overconfident agents as modeled by Daniel, Hirshleifer, and Subrahmanyam (1998). Such agents consider their private signals to be more precise than others' signals. The investors' overconfidence leads to underreaction in the short term and reversal in the long term. In their model, if a public event happens due to the firm being overvalued (undervalued) then there is positive (negative) abnormal return on the announcement day. Following the announcement, the returns have positive autocorrelation. Daniel, Hirshleifer, and Subrahmanyam (1998) cite stock issuance by firms as an example of a public event that is triggered by firms' overvaluation. Analyst upgrades (downgrades) are arguably also triggered by overvaluation (undervaluation) of stocks. This paper reports that the contemporaneous return related to the news that cite analysts is substantial. The return is almost 5% per month if the sentiment score of news article citing an analyst is extremely positive and -5% per month if the sentiment score is exteremly negative. By contrast, there is no outperformance from a long-short sentiment score-based trading

[6] From the bankruptcy court proceedings as recorded by Valukas (2010).

> Buffett spent the rest of Friday, March 28, 2008, reviewing Lehman's 10-K and noting problems with some of Lehman's assets. Buffet's concerns centered around Lehman's real estate and high yield investments, lending-related commitments derivatives and their related credit-market risk, Level III assets and Lehman's securitization activity.

strategy that uses news articles citing analyst opinion. This result suggests that the agents incorporate sentiment of analyst news into their beliefs about stock prices.

Investors could be boundedly rational as modeled by Hong and Stein (1999). Their model has two type of traders, those who react to news but not to price information, and those who do not react to news and only react to price information. Their model generates initial underreaction followed by overreaction. In their model, underreaction-based trading strategies in general, and momentum trading strategies in particular, are more profitable among smaller stocks than among larger stocks. The implication hinges upon the assumption that information diffuses slowly for smaller stocks. In this paper, the sentiment-based trading strategy is profitable among larger stocks. The profitability of a sentiment-based trading strategy for large stocks also indicates that the sentiment information from news diffuses slowly among large stocks as well.

Rational-inattention based explanation

Rational inattention treats attention as a scarce resource that an agent allocates over competing opportunities to deploy attention. Peng and Xiong (2006) model an agent's attention to be allocated over various information categories. The agent allocates more attention to economy-wide and sector-related information than to stock-specific information. As a result investors underreact to firm specific information. Arguably, news about large stocks contains more information about their respective sectors and the economy as a whole than news about small stocks. Therefore, agents should allocate significant attention to news about large stocks. Casual

empiricism supports the conjecture that earnings announcements of companies like Intel, Walmart, and Microsoft are much closely watched by investors than those of small stocks. This thinking would suggest that investors underreact to information pertaining to small stocks.

Nieuwerburgh and Veldkamp (2009) model agent's attention allocation over various investment choices. In their model, the agent allocates more attention to larger stocks than to smaller stocks. This implies that there will be more mispricing due to underreaction for smaller stocks than for larger stocks, perhaps with no mispricing for the largest stocks.

The two rational-inattention theories above suggest that there should be less underreaction for the largest stocks. In contrast to these theories, this paper finds that the sentiment-score based trading strategy is profitable for the largest decile of 3000 stocks.[7] Alternatively, if the agent was instead allocating attention over various information sources then the agent would pay more attention to analyst opinion. A sentiment-based trading strategy that trades on past sentiment of analyst recommendations does not generate any abnormal return, i.e. the market participants allocate more attention to articles mentioning analysts than to other news.

Market-friction based explanation

It is also possible that the stocks that are showing the underreaction phenomenon are the stocks that have high transaction costs as shown by Lesmond,

[7] The sentiment scores are observed with higher precision for larger firms than for smaller firms. It is possible that with better data on smaller firms, a sentiment-based trading strategy will be profitable for smaller firms as well.

Schill, and Zhou (2004). The trading strategy based on sentiment score is more profitable among more liquid stocks which typically have lower transaction costs. Profitability of a sentiment score based trading strategy for liquid stocks rules out the market-friction based explanation.

1.1.2 Nature of information in news articles

The findings in this paper also throw some light on the nature of information contained in news. Tetlock (2007) observes a short term reaction to pessimism in news articles, followed by a correction within next few days. This paper extends the findings of Tetlock (2007) to the longer term. In his interpretation, news is not informative. In this paper, a sustained reaction to news is observed over a long period of many months. The final month portfolio of both the sentiment score and the size-adjusted newscount strategy are trading on news articles that are at least five months old. The sustained reaction to news sentiment score and size-adjusted newscount indicates that news contains information.

Section 1.2 describes the data used in the study. Section 1.3 describes the methodology followed in this paper. Section 1.4 examines sentiment-score based trading strategies. Section 1.5 examines newscount based strategies. Section 1.6 reports the impact of liquidity on trading strategies. Section 1.7 sums up the implications of findings for future research and concludes.

1.2 Data

1.2.1 News data

The news data is obtained from the Thomson Reuters NewsScope dataset. It consists of all firm-specific news. Thomson Reuters provides firm-level sentiment information for all the firms mentioned in a news item. The sentiment is derived from the text of the news item. If a news item mentions multiple firms, each firm receives a different sentiment score based on the words used to describe the firm. The unit of analysis is a sentence rather than a word. Consider the following example from the news database.

> Analysts skeptical of Ford's better than expected earnings forecast.
>
> DETROIT, Jan 13 (Reuters) - Wall Street analysts gave a skeptical reception on Monday to Ford Motor Co.'s better-than-expected earnings forecast for this year, saying Ford's assumptions for flat prices and stable U.S. market share were too rosy.

The example article is four paragraph long, for the sake of brevity only the first paragraph is reproduced here. According to the Harvard IV Psycho-social dictionary each paragraph of this article has an equal number of positive and negative words. General Inquirer deems the article as neutral, however the Reuters' text processing engine judges the article as slightly negative. [8] Based on the reading the title or the first paragraph, the article seems to be negative. Let us consider an example that

[8] General Inquirer is a popular text processing engine which uses the Harvard IV Psycho-social dictionary to analyze the text.

describes the working of the text processing engine. Since the unit of analysis is a sentence, the working of the sentiment engine is demonstrated with one sentence. Figure 1.2 provides a schematic description of the same example.

BP gave analysts a negative surprise.

The sentiment engine has three major sequential processes: (1) pre-processing, (2) lexical and sentiment pattern identifier and (3) sentiment classifier. In the pre-processing stage the sentence is identified as composed of a subject and verb-phrase. In computational linguistics this is known as shallow parsing. The parts of speech of each word in the sentence are identified. Output from this stage is fed into the lexical and sentiment pattern identifier. Some parts of speech are more important for the purposes of communicating the tone of the sentence.In this case, the verb ('gave') and adjective ('negative') are important. In this same stage, the verb 'gave' is mapped to 'give'. Through these two processes, the subject is retained. The subject information is used for providing sentiment to the subject. It is also used to assign relevance. The features from the lexical and sentiment pattern identifier are fed into a three layer back-propagation neural network classifier with weight relaxation.

The text processing engine is described in detail in appendix 2.6. More details are available in Infonic (2008). The sentiment is reported as the probability of the article being positive, the probability of the article being negative and the probability of the article being neutral. The three probabilities sum to 1.

The fields that are used for construction of the sample are the following : Reuters instrument code (ticker and exchange), probability of the article being positive, probability of the article being negative, relevance of the article for the firm, "Item type" which indicates if the item was an "article" or an "alert", "Lcnt" (indicating the novelty of the item) and topic code (flags that indicate the industry of the news, analyst call, earnings, earnings forecast etc.). Figure 1.3 provides a screen shot of the raw database. In the example screen shot, the three news items are article, alert and article on Pilgrim's Pride, Ford and US Airways respectively. The news items are earnings announcement,research update and research updates respectively.

Sometimes more than one firm is mentioned in a news item. Thomson Reuters also provides relevance information for each firm named in the news. The relevance ranges from 0 to 1. A relevance value of 1 indicates that the news item is highly relevant to the firm. Some news items repeat information from a previous article. The Thomson Reuters database contains link count, which is the novelty of the news item. News item which has relevance of at least 0.35 and link count less than 2 are retained in the sample. Alerts are taken out. From the news database, monthly "newscount", average of the probability of being positive, negative and neutral at CUSIP level are obtained.

A measure of long term sentiment is constructed from monthly sentiment scores. The monthly sentiment scores are obtained by taking the average over all news items for a firm each month. A measure of long term sentiment is created by taking average of monthly scores over 3 months. Figure 1.4 shows the averaging

methodology. News comes at uneven intervals during a month. First the difference of the probability of positive article and negative article are averaged to obtained a monthly sentiment score. Then each month, a running average of past three months' monthly score is calculated. There are two reasons for choosing three months horizon for obtaining the long term sentiment. First, firms could strategically time the release of news. Graham, Harvey, and Rajgopal (2005) surveyed more than 400 finance executives and find that two thirds of respondents delay bad news to allow analysis and interpretation. Almost one third of respondents package bad news with other news. Monthly sentiment scores based on the past three months should help overcome this kind of packaging since at any time the long term sentiment contains all the information in past three months rather than just a point in time snapshot. Second, there is more coverage of firms during the earning season, with three month running window, each month long term sentiment score includes one earning season.

The empirical results in this paper show that long term sentiment score is a useful measure of information for the firms in the sample. Let us consider the example of Citi (Ticker symbol "C"). It is a large financial institution which had two periods of intense negative publicity during the sample period. First concerning CEO succession, and second concerning the financial crisis. Figure 1.5 shows the long term sentiment for Citi during the sample period. The bars show the monthly return. The correlation between long term sentiment and monthly return is 0.23. However, the long term sentiment measure does not take into account the cross-sectional variation in sentiment score. Very few financial firm had positive news during the financial crisis. Figure 1.6 shows Citi's long term sentiment when nor-

malized for cross sectional long term sentiment across firms. The normalized long term sentiment shows how much more negative Citi's relative long term sentiment surrounding the financial crisis was compared to earlier period of CEO succession. The correlation between normalized long term sentiment and monthly return is 0.34.

1.2.2 Financial data

The sample of firms consists of all US common stocks that are covered by Thomson Reuters NewsScope news services between 2003 and 2008. The firms with negative book value, as reported in the Compustat database, are excluded. Monthly stock returns are provided by Center for Security Prices (CRSP) and accounting data are provided by Compustat through the Wharton research data services (WRDS). WRDS also provides the Fama-French three factors as in Fama and French (1993) and the UMD factor. The book-to-market ratio is defined as Book Equity(BE) over Market Equity (ME). BE is the book equity for all fiscal year-ends in previous calendar year. Book equity is defined as Book Equity = Total Assets$-$[Total Liabilities$+$ Preferred Stock] + Deferred Taxes + Convertible Debt

as in Kayhan and Titman (2007). ME is the market capitalization at December of the previous year. In addition, market capitalization and return for each firm for every month are obtained. Firms are also sorted by market capitalization at the end of the previous month. Smallest firms are assigned to the lowest decile. Momentum deciles are based on last 6 months cumulative return. The first momentum decile has firms with worst cumulative return in past six months.

To measure liquidity at the individual stock level the Amihud illiquidity mea-

sure is calculated. The Amihud illiquidity measure is the absolute value of return for one day's dollar volume. It is calculated for all stocks at a monthly level by taking the average of

$$1000000 * abs(return)/(abs(price) * volume)$$

over all trading days of a month, as in Amihud (2002).

The sample consists of firms that are in the intersection of the WRDS-CRSP-Compustat database and Thomson Reuters news sentiment database. The two datasets are merged on CUSIP of the firm at year level. After merging these two databases, the sample has 587,719 stories over the entire sample period of which 41,776 are in 2003, 48,787 are in 2004, 79,083 are in 2005, 110,741 are in 2006, 135,619 are in 2007, and 171,913 are in 2008. The sample has 197,188 firm month observations or almost 3000 firms.

1.2.3 Summary statistics

Table 1.1 shows Thomson Reuters coverage aggregated at monthly level by each year of the sample. For the largest size decile the sample covers over 90% firms of the Compustat-CRSP intersection during the sample period. For smallest firms, the sample covers only 12% in year 2003, but 40% in year 2004. Thus, relative to the general population of US common stock with positive book value, the sample is biased towards larger firms.

The table 1.1 also shows the distribution of Thomson Reuters coverage by book

to market deciles and momentum deciles. For the largest B/M decile (value firms), the sample starts with only 35% of firms having any news, while the smallest B/M decile (growth firms), the sample in the year 2003 has 65% firm with some news. For the same period, there are fewer firms in extreme momentum deciles. In the year 2003, decile 1 and decile 10 have 43%, and 51% firms with some news, while decile 5 has 62% firms with some news. The pattern of more news for growth stocks than for value stocks and more news for moderate momentum stocks than for extreme momentum stocks also suggests that growth firms are on average bigger than value firms, and extreme momentum firms are smaller than moderate momentum stocks.

Table 1.2 shows average monthly return, size, newscount, standardized newscount, the probability of being positive, and the probability of being negative across size, book to market and momentum deciles.

The top panel of table 1.2 groups firms into 10 deciles sorted by market value of the firm. The average market cap of decile 1 is 21 million USD, while that of decile 10 is 261.6 billion USD. The smallest market cap decile has 0.15 news items per month, while the largest has 15.99 news items per month. Newscount is also monotonically increasing as one moves down the size deciles, indicating that large firms have more news than small firms.

The middle panel of table 1.2 groups firms sorted by increasing book to market. The table shows that value firms are smaller. The average size of growth firms (deciles 1, and 2) is 6.2 billion USD while the average size of value firms (deciles 9, and 10) is 1.9 billion USD. The lowest book to market firms get on average 3.29 newscount per month while value firms get only 2.05 news items. For the

same groups, the size adjusted news count -0.03 and 0.19 respectively. No pattern in sentiment stands out in either size or book to market tables. This indicates that value firms have less news than growth firms, but have more news when the newscount is adjusted for size.

The bottom panel of table 1.2 is sorted by momentum groups. Firms in decile 1 have the worst performance in past 6 months, while the firms in decile 10 have the best performance. The table indicates that momentum is pronounced among small stocks and stocks with high turnover. Deciles 1,2,9, and 10 have market cap of 781,2188, 3666, and 1755 million USD respectively. These deciles have monthly turnover of 24.5%, 18.4%,20.3% and 34.8% respectively. The median turnover across the 10 momentum deciles is 17.1%. The table also indicates that momentum stocks have higher size adjusted newscount. The average size-adjusted newscount in deciles 1,2,9, and 10 are 0.34,0.07, -0.05 and 0.04 respectively, while those of the firms in the middle (4,5, and 6) are negative. Stocks with positive momentum (deciles 8,9, and 10) have higher positive sentiment (0.32 for all three) compared to stocks with negative momentum (deciles 1,2, and 3) which have average positive sentiment of 0.22,0.26, and 0.27. Similarly, stocks with positive momentum (deciles 8,9, and 10) have lower negative sentiment(0.24 for all three) compared to stocks with negative momentum (deciles 1,2, and 3) which have average negative sentiment of 0.35,0.32, and 0.30.

The sentiment score is measured as the difference between the probability of the article being positive and the article being negative. Table 1.3 shows that with increasing sentiment score, the contemporaneous return increases (from a -1.94%

monthly return for decile 1 to a 3.54% monthly return for decile 10). This lends economic validity to the sentiment scores. The fact that the average sentpos is increasing across groups and average sentneg is decreasing across groups indicates that differencing the two probabilities does not lead to destruction of much information. The table also shows that smaller firms are more likely to receive extreme sentiment news in a month. The market cap for deciles 1,2,9, and 10 are 1.6 billion,5.1 billion,4.1 billion, and 2.6 billion USD respectively, while that of decile 4,5, and 6 are 14.4 billion, 13.2 billion, and 10.1 billion USD respectively.

Table 1.4 sorts firms by increasing standardized newscount. The table shows average contemporaneous return, size, book-to-market, and sentiment scores. The table indicates that with increasing standardized newscount, the probability of the news being negative increases. The probabilities of having positive news are 0.31,0.32, and 0.32 for deciles 1,2, and 3, while they are 0.30,0.29, and 0.26 for deciles 8,9 and 10. The table also shows that turnover is higher for high and low standardized newscount, 0.19,0.17, 0.24, and 0.34 for deciles 1,2,8, and 10 respectively. Contemporaneous returns increase with increased standardized newscount up to a point and then decrease with increasing standardized newscount (inverted U shape).

1.3 Methodology

This paper uses two methodologies: (1) regressions of returns on firm level characteristics and (2) portfolio regressions on Fama-French factors and UMD. As a

robustness checks Fama-MacBeth procedure is also applied to all specifications with firm level regressions.

1.3.1 Firm level characteristic regression

The return is estimated as a function of firm characteristics with date fixed effects. Applying date fixed effect has the same effect as applying β of one for all stocks. In this specification, long term sentiment is also a firm characteristic. In all specifications, the normalized long term sentiment is used unless specified otherwise. The controls are book to market, firm size from the previous month, and previous three month returns.[9]

$$r_{it} = \beta_t \times d_t + \beta_j \times X_{jit} + \epsilon_{it} \tag{1.1}$$

One of the concerns with news data is that there are very few news items for small stocks. Fewer observations lead to heteroskedasticity of a known form, where the estimates are less precise for smaller firms. Firm level regression (1.1) allows the flexibility to apply different weighting schemes and thereby control for heteroskedasticity. A weighting scheme of $\frac{1}{n}$, where n is the number of news article in the previous month puts more weight on smaller firms. Weighting scheme of n, where n is the number of news article in the previous month puts more weight on larger firms.

Results from regression (1.1) indicate that long term sentiment or past sentiment has predictive ability for returns but it does not indicate whether long term

[9] Results are qualitatively similar for six months and twelve month returns.

sentiment is driving returns' predictability or long term sentiment itself is being driven by past return as well. To differentiate between the two possible drivers of return predictability, an alternative specification is applied. First, the returns are regressed on all the firm level characteristics except long term sentiment. The residuals are regressed on long term sentiment. If predictability of returns from long term sentiment is largely driven by long term sentiment, then the regression coefficient on long term sentiment should be similar to the one from estimate (1.1). Specification (1.2) formally expresses this specification, where LTS denote long term sentiment.

$$\begin{aligned} r_{it} &= \beta_t \times d_t + \beta_j \times X_{jit} + \epsilon_{it} \\ \epsilon_{it} &= \beta_{\text{LTS}} \times \text{LTS} + \eta_{it} \end{aligned} \tag{1.2}$$

Similarly, returns could be first regressed on long term sentiment and other firm characteristics except for past returns. Residuals from this regression can be regressed on past returns. If the returns predictability is driven by past return then the regression coefficient on returns of months 1 ,2 ... 6 should not be different from specification (1.1).

1.3.2 Portfolio regression

Portfolios based on news sentiment score and size-adjusted newscount are constructed. For the news sentiment specification, first the stocks are sorted on past three month sentiment. The sorting has the same effect as normalizing sentiment, since it is a cross sectional sorting.

For the size adjusted newscount portfolio, first the size-adjusted newscount

is obtained. Each month all the firms are sorted on the basis of previous month's market capitalization into deciles. For each size decile, the average newscount and standard deviation of newscount are calculated. The Z score is assigned as,

$$Z_i = \frac{\text{firm}_i \text{ newscount} - \text{average newscount of firms in firm i's decile}}{\text{standard deviation of newscount of firms in firm i's decile}} \tag{1.3}$$

Figure 1.4 shows the timeline of portfolio formation and holding. Each month, three month average sentiment score and standardized newscount are calculated. Firms get more news around earnings release. Having a window of 3 months ensures smooths the number of news in each formation period. They are called past sentiment score (or long term sentiment score) and past size-adjusted newscount. All the stocks are sorted on the basis of the past sentiment score into deciles. The sentiment based trading strategy is to take a long position in stocks in the top two deciles of past sentiment score, and take a short position in stocks in the bottom two deciles of past sentiment score. The newscount based trading strategy is to take a long position in stocks in the middle three deciles (4,5,6) of past size adjusted newscount, and take a short position in stocks in the top and bottom two deciles (1,2,9, and 10) of past size adjusted newscount.[10]

It is possible that firms strategically time the release of news. Graham, Harvey, and Rajgopal (2005) surveyed more than 400 finance executives and found that two thirds of respondents delay bad news to allow analysis and interpretation. Almost one third of respondents package bad news with other news.

[10] The results are robust to sorting in quintiles and fifteen groups as well.

Portfolio performance is examined along two dimensions. First, the correlation of portfolio return with momentum is examined. Second, whether the portfolios can generate significant return in excess of Fama-French three factors and UMD.

Market participants may underreact to information as a result of not trading of on information of sufficiently low economic significance to justify the transaction cost in carrying out a trade. To understand the economic significance, the portfolio returns are examined by size deciles. The stocks are first sorted into size decile and the portfolio returns are then examined.

The profitability of the trading strategy could be due to the momentum phenomenon. Following Daniel and Titman (2007), the stocks are first sorted into momentum deciles. Subsequently the trading strategies are examined whether they are still profitable.

For further robustness, the stocks are sorted by the Amihud illiquidity measure and the profitability of portfolio strategy is examined by illiquidity deciles.

1.4 Sentiment of news

The table 1.5 shows the average monthly return across deciles sorted by past three month average of sentiment scores. Decile 1 has most negative average sentiment score while decile 10 has most positive average sentiment score. The table also shows formation standardized newscount, size, turnover, newscount, and future evolution of sentiment. While the average return of decile 10 is higher than decile 1 (87 basis points per month versus 50 basis points), the difference is not

statistically significant. Furthermore, the size varies in a U shape pattern across the deciles, indicating that smaller firms have extreme sentiment news. Also firms with no news have average market cap of 277 million, smaller than any other sentiment group. Firms with no news also have much lower turnover than any other sentiment group. Given the dispersion in firm size across sentiment portfolios, it is reasonable to examine long-short sentiment portfolios within size categories.[11]

The table 1.6 shows the estimates for specification (1.1) for three different weighting schemes. Firm specific return are controlled for date fixed effect, book to market, size and previous three month return. Model 1 is equally weighted, assigning more weight to larger firms. A percentage point increase in long term sentiment score translates to an 11 basis points increase in returns. Model 2 assigns more weight to firms with smaller number of news items in the previous month. For smaller firms, a percentage point increase in the long term sentiment amounts to a 76 basis points increase in return. Model 3 assigns more weight to firms with large number of news items in the previous month. This scheme assigns far more weight to large firms. Since the number of news item increases at a faster rate than the size of firms, this assigns more weight to larger firms than a value weighted scheme. This weighting scheme indicates that one point increase in the long term sentiment amounts to 8 basis points increase in return. All, the estimates show that long term sentiment predicts future return.

Table 1.6 shows that long term sentiment predicts returns even when controlled

[11] Since smaller firms have lesser amount of news per month, monthly sentiment score are observed with lesser precision compared to larger firms. This is a limitation of data and to this extent conclusions about smaller firm groups suffer from not having precise measure of monthly sentiment.

for past six month returns (momentum). The estimation is carried out as specified in specification (1.1). First returns are controlled for market, firm size, book-to-market, and past six months return. If the long term sentiment has any predictive ability left, that would indicate that the predictive ability of long term sentiment is independent of past returns. The estimates for the first stage are shown in the top panel. The top panel estimates show that there is predictive ability in returns in the sample period. Also there is reversal in the first month. The residuals from the first stage are regressed on long term sentiment. Estimates are presented in model 1 of the second panel. A one point increase in long term sentiment is associated with a 20 basis point increase in monthly return. Model 2 of the second panel presents estimates for past six months returns, controlling for past sentiment. The estimates show that past six returns have no predictive ability controlling for long term sentiment (or past sentiment). The table also shows that the well known reversal in the first month actually changes sign when controlling for long term sentiment.

The table 1.8 shows the alpha and factor loadings from the long-short sentiment strategy by size deciles. The alpha is calculated in excess of the Fama-French 3 factors and UMD. Almost all portfolios have statistically significant coefficient on the UMD factor. For the 10 size long short portfolios, the average adjusted R square is only 12.4% as opposed to 24% when the UMD factor is included. Three out of ten size portfolios have economically and statistically significant alphas(size group 3, 8, and 10 have alpha of 175, 38, and 34 basis points respectively). Note that two out of these three size portfolios with economically significant returns are top three

size deciles.

Figure 1.7 shows the performance of the long-short sentiment portfolio during the sample period. As a zero-value long-short portfolio, it is long one dollar and short one dollar. The portfolio begins with zero dollars and the portfolio return every month is added to the existing amount. The portfolio has a cumulative return of 22 cents. The figure shows hedged and unhedged long-short portfolio as well as excess return on the market portfolio. The hedged portfolio return is obtained by taking out the exposure to the Fama-French three factors and UMD.

Figure 1.8 shows long-short momentum portfolio (UMD) and long-short sentiment score strategy applied on the largest two deciles of firms. It shows that the UMD and the sentiment portfolio have highly correlated returns. The correlation is higher in the second part of the sample, when UMD has positive returns.

This paper claims that the sentiment-score based strategy is related to the momentum phenomenon. To demonstrate it, table 1.9 shows alphas in excess of the Fama French three factors and UMD from the sentiment score based long-short trading strategy when deployed across ten momentum deciles. Stocks are first sorted into deciles based on past 6 months return. Within each momentum decile, all stocks are sorted by sentiment. A long position is taken on positive sentiment score stocks (deciles 9, and 10) and short position is taken on negative sentiment stocks (deciles 1, and 2). The portfolio is held for five months. Each month the portfolio has five vinatges, one from each month. Five out of ten momentum groups have economically and statistically significant alphas(74, 58,72, 40, and 43 basis points in momentum deciles 2,3,4,5, and 7 respectively). As one would have inferred from size-sorted

long-short portfolios, momentum groups 3,4,5,6 and 7 are the larger stocks in the sample.

The regression coefficients on UMD demonstrate the relationship of long-short sentiment score strategy with momentum. Given that these are long-short strategies within momentum groups, the coefficients on UMD factor are expected to be zero. In fact, the UMD factor loading is positive for all ten groups, and is statistically significant for eight out of ten momentum groups. Interestingly enough, the factor loading is positive even for negative momentum groups 1,2,3,4, and 5.

Table 1.10 shows contemporaneous returns for portfolios sorted on sentiment scores calculated based on the subset of news articles about analyst coverage. The monthly sentiment score of such stories is obtained. The table shows that analyst tend to cover large firms. Average market capitalization of all ten sentiment deciles are in excess of 10 billion dollars. The contemporaneous return in each of the ten decile portfolios is economically significant. For an extremely negative sentiment score (decile 1) the stock is down 5% in the month, and is up 5% for an extremely positive sentiment score (decile 10). A long-short stragety based on sentiment score does not generate any alpha. Since all the alpha is in the current month and no alpha is the following months, this indicates that the market participants pay attention to sentiment information of news that attribtes analyst.

1.5 Normal volume of news

Table 1.11 shows monthly equally-weighted and value-weighted returns for portfolios sorted by increasing standardized newscount in the formation period. Each month, we look back past three months and sort based on size-adjusted newscount. The portfolios are held for five months. The table shows an inverted U-shaped pattern for equally weighted returns. The equally-weighted monthly returns for deciles 4,5, and 6 are 1.18%, 0.94%, and 0.80% respectively, while the returns of deciles 1,2,9 and 10 are 0.58%, 0.69%, 0.70%, and 0.59% respectively. For value-weighted monthly returns, no such pattern exists. This indicates that if there is any alpha for this particular strategy, it comes from smaller stocks in each group. The turnover continues to be higher for stocks with unusually low and unusually high amounts of news.

Table 1.12 shows the alpha on long-only portfolios across historical size-adjusted newscount deciles. The Newcount portfolio for decile 1 has the lowest amount of standardized news in past 3 months. Portfolio 4,5, and 6 are the portfolio of firms with usual amounts of news in the formation period. The returns are calculated in excess of prevailing risk free rate. The table shows portfolio alphas of 49, 38, and 44 basis points per month in excess of the Fama-French three factors and UMD for portfolios 4,5, and 6 respectively. All the ten decile portfolios co-move with the SMB factor. Also all the ten decile portfolios have negative loading on UMD. Portfolios with historically higher size adjusted news tend to have higher market beta. The alphas of the portfolios with usual amounts of historical newscount are economically

and statistically significant.

Figure 1.9 shows the performance of the long-short size-adjusted newscount portfolio during the sample period. As a zero-value long-short portfolio, it is long one dollar and short one dollar. The portfolio begins with zero dollars and the portfolio return every month is added to the existing amount. The portfolio finishes at 29 cents. The chart has hedged and unhedged long short portfolio as well as excess market return.

Table 1.13 describes alphas from the long-short newscount-based strategy across momentum groups. Stocks are first sorted in momentum deciles based on past six months returns. Within each momentum decile, further sorting is done based on increasing historical standardized newscount. The long-short portfolios are formed by taking a long position in deciles 4,5, and 6 and short position in deciles 1,2,9, and 10. The long-short portfolios are held for 5 months after formation. In each month, there are 5 vintages and the portfolio return is calculated by averaging over these five vintages. Of the ten momentum long-short portfolios, five have statistically and economically significant portfolio alphas(53 basis points, 28 basis points, 29 basis points, 37 basis points and 80 basis points) in excess of the Fama-French factors and UMD. Among five of the portfolios with significant alpha, only one loads significantly on SMB. These are long-short portfolios within momentum group. The coefficient on the UMD factor is expected to be zero. By contrast, moving down the momentum deciles, one notices that the portfolios 1,2,3,4, and 5 have positive co-movement with the UMD factor while portfolio 9, and 10 have negative and statistically significant coefficient on UMD. This suggests that when

stocks have normal amounts of news and had poor past performance, they actually do better than their momentum peers. Also if stocks have an abnormally high or low amount of news and had good past performance, they undergo reversal in subsequent months.

It is possible that these results are just another artifact of the relationship between no news and positive alpha as reported in Chan (2003). I examine this conjecture by creating portfolios of stocks with no news across momentum groups. I regress the portfolio return in excess of risk free rate for those portfolios on Fama-French 3 factors and UMD. The regression shows that no news is good news for stocks which have been doing well. The alphas for momentum portfolios 8, 9, and 10 are 68, 101, and 80 basis points respectively. No-news portfolio have negative coefficient on UMD for all momentum deciles. The coefficient for no news portfolio are almost 1 for the SMB factor, indicating no news is far more likely for smaller stocks. The regression also shows that the no-news portfolio and long short news-count portfolio discussed earlier are different portfolios.

1.6 Role of Illiquidity

The Amihud illiquidity measure (Amihud 2002), is measured for all stocks at monthly level by using the formula 1000000*abs(ret)/(abs(prc)*vol) for all stocks in the CRSP daily stock file. All the stocks are sorted in decile by this illiquidity measure. Stocks in decile 10 are most illiquid stocks in the sample.

Table 1.14 shows the alphas from a long-short sentiment strategy by illiquidity

group. Six out of ten liquidity-sorted portfolios have economically and statistically significant alphas. The alpha in illiquid group tends to be higher than the liquid groups, indicating more liquid stocks have lesser under-reaction than the illiquid stocks.

Table 1.15 shows the alphas from the long-short newscount strategy by illiquidity decile. Decile 1 has the most liquid stocks while decile 10 has the most illiquid stocks in the sample. The decile with most illiquid stocks has alpha of 69 basis points per months in excess of the Fama-French three factors and UMD. Most liquid stocks have no alpha. In fact the alpha for second most liquid stocks is -30 basis points. Clearly liquidity plays a role for the newscount-based strategy.

1.7 Implications and conclusion

1.7.1 Implications

Past sentiment score (or long term sentiment score) predicts future returns. Controlling for the past sentiment score, the predictive ability of past returns (to predict future returns) disappears. The sentiment score is an example of a quantitative measure of qualitative information. The evidence presented in this paper is consistent with the hypothesis that slow absorption of qualitative information could be driving the momentum phenomenon. Since both models of rational inattention and behaviorally impaired investors predict such underreaction to be restricted to small stocks. The findings presented here do not fit either models of rational inattention or the models of behaviorally impaired agents. The findings do suggest that

information comes into prices slowly, which is consistent with the model where some agents have informational advantage and choose to trade like a monopolist.

The Trading strategy that uses past sentiment scores generates portfolios which are highly correlated with momentum and profitable even when controlling for the Fama-French three factors and UMD. Interestingly, the sentiment score based trading strategy is profitable among the largest size decile. This does not seem consistent with rational inattention based explanations of momentum since these models restrict underreaction to small stocks. The results of this paper are consistent with the hypothesis that investors pay more attention to news about analysts than to other news articles. This observation is somewhat at odds with popular behavioral models of investor inattention, since in such models there is underreaction to analyst news as well. It, however, fits with the idea that investors process hard information and soft information differently. Analyst upgrades and downgrades are closer to hard information than soft information. Sentiment score of all the articles about a firm, on the other hand, is a measure of soft information. It is also possible that models that treat different sources of information differently might be able to explain these findings better. For future studies, it will be interesting to separate the two possible explanations - (1) differential processing of hard and soft information, and (2) differential processing of information from different sources.

Large firms have more news items than smaller firms. This suggests that the number of news items should be adjusted for the size of the firm. Adjusting the newscount for size, one observes that abnormally high or abnormally low size-adjusted newscount is associated with poor future returns. The profitability of the

size-adjusted newscount based trading strategy suggests that the amount of news plays a role in investor behavior beyond the role noted by (Barber and Odean 2008). Barber and Odean (2008) note that stocks that are highly covered in news are bought by retail investors who tend to have poor information. That might explain why shorting stocks with high size adjusted newscount might be a profitable strategy. But it does not explain why shorting stocks that have *low* size adjusted newscount is also profitable. Therefore, the newscount based strategy needs to be studied in some more detail.

1.7.2 Conclusion

This paper examines two trading strategies based on news. There are three main results.

First, there is the relationship between a long term sentiment-score based trading strategy and UMD. The trading strategy which is short stocks with negative past sentiment score and long stocks with negative past sentiment score is profitable among the largest and more liquid stocks of the sample. The portfolio resulting from the long-short sentiment score based trading strategy is also highly correlated with UMD factor. The findings suggest that underreaction to sentiment information is related to the momentum phenomenon. It also suggests that the market participants process soft information differently from hard information.

Second, there is a linkage between size-adjusted newscount and future returns. Stocks with abnormally high and abnormally low size-adjusted newscount underperform stocks with normal newscount. Merton (1987) suggests that stocks with high

recognition have low risk-premium since they have lower information risk. Barber and Odean (2008) suggest that stocks with abnormally high newscount are bought primarily by retail investors. Retail investors sell stocks that are profitable to hold onto. Both papers suggest that being short on stocks with high size adjusted newscount should be profitable. However, it is also profitable to be short stocks that have abnormally *low* amount of size-adjusted newscount.

Third, the findings suggest that news articles contain information. Past news article sentiment score and size-adjusted newscount predict future performance. This suggests news contains information.

1.8 Tables and figures

Size Decile	2003	2004	2005	2006	2007	2008
1	11%	40%	56%	83%	91%	97%
2	22%	57%	68%	85%	94%	99%
3	28%	64%	82%	90%	95%	99%
4	46%	75%	82%	91%	95%	98%
5	55%	76%	87%	95%	98%	99%
6	65%	83%	89%	94%	98%	100%
7	72%	80%	87%	93%	97%	100%
8	76%	84%	88%	96%	98%	100%
9	88%	91%	94%	98%	99%	100%
10	95%	96%	97%	100%	100%	100%
B/M Decile	2003	2004	2005	2006	2007	2008
1	66%	77%	84%	93%	97%	98%
2	67%	79%	85%	94%	97%	100%
3	69%	80%	87%	96%	98%	100%
4	64%	81%	87%	93%	98%	100%
5	64%	76%	85%	95%	97%	99%
6	59%	81%	85%	92%	97%	99%
7	58%	77%	84%	92%	97%	99%
8	52%	73%	82%	93%	97%	99%
9	46%	66%	78%	92%	95%	99%
10	35%	62%	74%	87%	94%	99%
Momentum Decile	2003	2004	2005	2006	2007	2008
1	43%	63%	74%	90%	93%	99%
2	57%	72%	82%	93%	97%	99%
3	62%	75%	84%	94%	98%	100%
4	62%	78%	85%	93%	97%	100%
5	62%	79%	86%	93%	97%	100%
6	63%	79%	86%	93%	97%	99%
7	61%	78%	86%	93%	97%	99%
8	58%	78%	85%	94%	98%	99%
9	53%	75%	83%	92%	96%	99%
10	51%	72%	81%	91%	95%	99%

Tab. 1.1: **Coverage by Size, Book to Market and Momentum:** The top panel shows coverage of CRSP-Compustat firms by Thomson Reuters. Each year the number of firms that are covered by Thomson Reuters are expressed as percentage of total number of firms in each size decile. Decile 1 has the smallest firms, while decile 10 has the largest firms. The Middle panel shows coverage by book-to-market deciles, expressed as percentage of total number of firms in each B/M decile. Decile 1 has growth firms, while decile 10 has value firms. The lower panel shows coverage by momentum decile, expressed as percentage of total number of firms in each momentum decile. Decile 1 has firms with extremely negative returns in past six months, while decile 10 firms with extremely positive returns in past six months.

Size decile	Return	Size	Turnover	Newscount	SNewscount	Sentpos	Sentneg
1	3.30%	21	17.80%	0.15	-0.06	0.29	0.21
2	2.02%	53	12.20%	0.25	-0.01	0.26	0.25
3	1.10%	101	11.20%	0.34	-0.03	0.28	0.29
4	1.15%	182	14.80%	0.57	0.01	0.27	0.30
5	0.81%	290	18.50%	0.72	0.00	0.28	0.31
6	0.69%	483	22.00%	0.97	0.01	0.29	0.29
7	0.72%	818	24.20%	1.31	0.00	0.30	0.28
8	0.57%	1456	26.40%	2.00	0.00	0.31	0.27
9	0.62%	3095	23.30%	3.48	0.01	0.31	0.26
10	0.45%	26157	20.00%	15.99	0.05	0.29	0.25
B/M decile	Return	Size	Turnover	Newscount	SNewscount	Sentpos	Sentneg
1	0.43%	6918	23.80%	3.29	-0.03	0.29	0.27
2	0.51%	5415	21.70%	3.31	-0.03	0.30	0.26
3	0.68%	4734	20.50%	3.04	-0.04	0.30	0.26
4	0.74%	4223	18.80%	3.07	-0.04	0.30	0.26
5	0.90%	2785	18.00%	2.64	-0.05	0.31	0.26
6	0.83%	3257	17.60%	3.1	0.00	0.29	0.27
7	0.68%	2405	17.30%	2.51	-0.04	0.29	0.27
8	0.84%	2467	16.80%	2.51	0.00	0.29	0.28
9	1.15%	2505	16.40%	2.78	0.06	0.29	0.28
10	1.29%	1254	15.50%	2.05	0.19	0.26	0.31
Momentum decile	Return	Size	Turnover	Newscount	SNewscount	Sentpos	Sentneg
1	-9.10%	781	24.50%	1.9	0.34	0.22	0.35
2	-4.13%	2188	18.40%	2.3	0.07	0.26	0.32
3	-2.30%	3290	16.60%	2.7	0.00	0.27	0.30
4	-1.12%	4945	15.40%	3.2	-0.04	0.29	0.28
5	0.04%	5226	15.20%	3.4	-0.08	0.30	0.27
6	1.04%	5621	15.10%	3.6	-0.09	0.30	0.26
7	2.12%	5425	15.80%	3.5	-0.09	0.32	0.25
8	3.37%	5022	17.60%	3.3	-0.07	0.32	0.24
9	5.60%	3666	20.30%	2.7	-0.05	0.32	0.24
10	11.85%	1755	34.80%	1.8	0.04	0.32	0.24

Tab. 1.2: **Summary statistics by Size, Book to Market, and Momentum:** The top panel shows the average monthly return, market capitalization (in million USD), turnover, newscount, standardized newscount, probability of being positive and the probability of being negative by size deciles. Decile 1 is small firms. The middle panel shows the same variables by book-to-market deciles. Decile 1 is growth firms while decile 10 has value firms. The bottom panel shows the same variables by momentum deciles. Decile 1 has firms with extremely negative returns in past six months, while decile 10 firms with extremely positive returns in past six months.

SentGrp	Sent		Return		Size		Newscount		Snewscount		Sentpos		Sentneg	
1	-0.65	0.01	-1.94%	0.77%	1557	86	2.06	0.15	0.42	0.02	0.08	0.00	0.73	0.01
2	-0.37	0.02	-1.91%	0.73%	5060	555	6.42	0.72	0.75	0.03	0.16	0.00	0.53	0.02
3	-0.19	0.02	-0.37%	0.75%	10939	746	10.39	0.86	0.85	0.02	0.19	0.00	0.38	0.02
4	-0.07	0.01	-0.01%	0.65%	14386	584	11.01	0.66	0.78	0.03	0.21	0.01	0.28	0.02
5	0.01	0.01	0.84%	0.63%	13153	637	8.95	0.51	0.64	0.03	0.21	0.01	0.20	0.01
6	0.08	0.01	1.33%	0.68%	10106	489	6.96	0.4	0.56	0.02	0.25	0.01	0.17	0.01
7	0.15	0.01	1.73%	0.65%	8078	474	5.69	0.37	0.55	0.03	0.30	0.01	0.15	0.01
8	0.24	0.01	2.33%	0.70%	6018	354	4.63	0.29	0.50	0.03	0.37	0.01	0.13	0.01
9	0.38	0.01	3.25%	0.68%	4096	170	3.28	0.14	0.37	0.02	0.48	0.01	0.10	0.00
10	0.65	0.01	3.54%	0.63%	2608	81	1.91	0.06	0.14	0.02	0.70	0.01	0.05	0.00

Tab. 1.3: **Summary statistics across sentiment deciles:** This table shows average sentiment, return, market capitalization (in million USD), newscount, standardized newscount, average probability of being positive, and the average probability of being negative by sentiment group. The sentiment scores are obtained by taking the difference of sentpos and sentneg. It also shows standard errors for each of these variables.

Newscount Group	snewscount		Return		Size		Turnover		Newscount		Sentpos		Sentneg	
1	-0.72	0.01	0.23%	0.57%	3376	165	0.19	0.01	0.50	0.10	0.31	0.01	0.23	0.01
2	-0.59	0.01	0.44%	0.61%	2138	119	0.17	0.00	0.49	0.06	0.32	0.01	0.23	0.01
3	-0.51	0.01	0.42%	0.67%	1563	136	0.15	0.00	0.42	0.05	0.32	0.01	0.23	0.01
4	-0.43	0.01	0.87%	0.69%	1835	190	0.14	0.01	0.69	0.08	0.32	0.01	0.23	0.01
5	-0.35	0.01	1.06%	0.76%	2242	266	0.15	0.01	1.13	0.15	0.31	0.01	0.25	0.01
6	-0.25	0.01	1.19%	0.82%	2043	216	0.15	0.01	1.22	0.15	0.30	0.01	0.25	0.01
7	-0.11	0.01	0.63%	0.73%	3937	320	0.19	0.01	2.64	0.20	0.30	0.01	0.27	0.01
8	0.17	0.02	0.67%	0.64%	4766	221	0.23	0.01	4.12	0.18	0.30	0.01	0.27	0.01
9	0.68	0.02	0.90%	0.72%	5654	189	0.24	0.01	5.85	0.32	0.29	0.01	0.28	0.01
10	2.36	0.01	1.35%	0.86%	13072	236	0.34	0.01	13.28	0.71	0.26	0.01	0.28	0.01

Tab. 1.4: **Summary statistics across standardized-newscount deciles:** This table shows average standardized newscount, return, market capitalization (in million USD), turnover, newscount, average probability of being positive, and the average probability of being negative by standardized newscount groups. It also show standard error for each of these variables.

Past Sentiment Group	FormSent		FormScount		Return		Size		Turnover		Newscount		SentPos		Sentneg	
No news			-0.43	0.01	1.09%	0.76%	277	13	0.13	0.01	0.31	0.02	0.29	0.01	0.25	0.01
1	-0.60	0.01	-0.03	0.01	0.50%	0.78%	646	26	0.17	0.00	1.03	0.07	0.26	0.00	0.35	0.01
2	-0.33	0.02	0.24	0.02	0.48%	0.76%	2341	272	0.24	0.01	3.06	0.40	0.26	0.00	0.34	0.01
3	-0.19	0.01	0.34	0.01	0.61%	0.71%	5728	563	0.23	0.01	5.91	0.74	0.27	0.00	0.31	0.01
4	-0.08	0.01	0.33	0.01	0.52%	0.68%	9242	299	0.22	0.00	6.77	0.44	0.27	0.01	0.28	0.01
5	0.00	0.01	0.28	0.02	0.48%	0.66%	11159	226	0.22	0.01	6.84	0.27	0.28	0.01	0.26	0.01
6	0.08	0.01	0.12	0.02	0.51%	0.66%	7533	330	0.20	0.01	4.20	0.15	0.30	0.01	0.25	0.01
7	0.14	0.00	0.10	0.02	0.46%	0.63%	5345	226	0.20	0.00	3.16	0.13	0.31	0.01	0.24	0.01
8	0.22	0.00	0.00	0.02	0.72%	0.64%	3674	203	0.20	0.00	2.13	0.08	0.32	0.01	0.23	0.01
9	0.33	0.01	-0.05	0.01	0.58%	0.65%	2791	88	0.22	0.01	1.85	0.07	0.34	0.01	0.23	0.01
10	0.55	0.01	-0.15	0.01	0.87%	0.64%	1618	48	0.20	0.00	1.22	0.05	0.35	0.01	0.23	0.01

Tab. 1.5: **Summary statistics across formation sentiment deciles:** This table shows average formation sentiment,formation standardized newscount,equally weighted return, market capitalization (in million USD), turnover, newscount, average probability of being positive, and the average probability of being negative by formation sentiment deciles. The portfolios are formed by sorting on the average of sentiment-score (probability of being positive - Probability of being negative) and held for overlapping five months. The average of all the reported variables are taken over three months. The table also shows standard error for each of these variables.

	Model 1	Model 2	Model 3
Long term sentiment	0.0011 ***	0.0076 ***	0.0008 ***
Book to Market	Y	Y	Y
Firm Size	Y	Y	Y
$Return_{-1}$	Y	Y	Y
$Return_{-2}$	Y	Y	Y
$Return_{-3}$	Y	Y	Y
Fixed Effect	Y	Y	Y
Weight	No	$\frac{1}{n}$	n

Tab. 1.6: **Firm level regression:** This table shows the estimate of β_{LTS} from the specification. $r_{it} = \beta_t \times d_t + \beta_j \times X_{jit} + \epsilon_{it}$ The three models are different from one another with respect to the weighing scheme. Model 1 is equally weighted, model 2 assigns more weight to firms with smaller number of news items in the previous month, and model 3 assigns more weight to firms with large number of news items in the previous month. The estimates show that long term sentiment (or past sentiment score) predicts future return.
***: 1%, **: 5%, *: 10%,†: 15%.

First stage regression

Long term sentiment	N
log(Book to Market)	Y
log(Firm Size)	Y
$Return_{-1}$	-.01261 ***
$Return_{-2}$	0.0060 ***
$Return_{-3}$	0.0052 ***
$Return_{-4}$	0.0003
$Return_{-5}$	0.0032
$Return_{-6}$	0.0187 ***
Fixed Effect	Y

Second stage regression

	Model 1	Model 2
Long term sentiment	0.0020 ***	Y,N
$Return_{-1}$	Y,N	0.028 ***
$Return_{-2}$	Y,N	0.0024
$Return_{-3}$	Y,N	0.0008
$Return_{-4}$	Y,N	-.0009
$Return_{-5}$	Y,N	0.0043
$Return_{-6}$	Y,N	0.0043
Fixed Effect	Y,N	Y,N

Tab. 1.7: **Firm level regression:** This table shows the estimate of β_{LTS} controlling for past six month returns. Returns are first regressed on control variables, including past six month returns as follows. $r_{it} = \beta_t \times d_t + \beta_j \times X_{jit} + \epsilon_{it}$ The estimate for the first stage are shown in the top panel. The top panel estimates show that past returns predict future return. Also, there is reversal in the first month. The residuals from the first stage are regressed on long term sentiment. Estimates are presented in model 1 of the second panel. Model 2 of the second panel presents estimates for past six months return, controlling for past sentiment. The estimates show that past six returns have no predictive ability after controlling for long term sentiment (or past sentiment).
***: 1%, **: 5%, *: 10%,†: 15%.

Size Group	Alpha			Mktrf			HML			SMB			UMD		
1	-1.08%	1.01%		0.20	0.29		0.42	0.55		0.41	0.48		0.52	0.29	**
2	-0.17%	0.52%		-0.21	0.15		0.18	0.28		-0.04	0.25		0.43	0.15	***
3	1.75%	0.42%	***	-0.37	0.12	***	-0.42	0.23	*	0.07	0.20		0.19	0.12	*
4	0.32%	0.35%		-0.12	0.10		0.12	0.19		-0.11	0.17		0.24	0.10	***
5	0.07%	0.31%		-0.23	0.09	***	-0.23	0.17		-0.17	0.15		0.03	0.09	
6	0.12%	0.20%		-0.05	0.06		-0.27	0.11		-0.16	0.10	*	0.14	0.06	**
7	0.03%	0.20%		0.12	0.06	**	0.06	0.11		-0.23	0.10	**	0.27	0.06	***
8	0.38%	0.22%	*	-0.05	0.06		-0.31	0.12		-0.22	0.10	**	0.33	0.06	***
9	0.20%	0.20%		0.16	0.06	***	-0.25	0.11		-0.17	0.10	**	0.33	0.06	***
10	0.34%	0.19%	*	0.07	0.05		-0.39	0.10	***	-0.07	0.09		0.24	0.05	***

Tab. 1.8: **Long short sentiment score portfolio across size deciles:** This table shows the coefficient, standard error and p-value from regressing monthly long-short sentiment equally weighted portfolio returns on Fama-French three factors and UMD. The decile portfolios are formed by sorting on past 3 month average sentiment within each size decile. The portfolios are long groups 9, and 10 and short groups 1, and 2. The portfolios are held for five months, with each month having five vintages. The returns are calculate by averaging over these vintages each month.
***: 1%, **: 5%, *: 10%,†: 15%.

Momentum Group	Alpha			Mktrf			HML			SMB			UMD		
1	-0.62%	0.50%		-0.15	0.14		-0.17	0.27		0.09	0.24		0.36	0.14	***
2	0.74%	0.31%	**	-0.32	0.09	***	-0.53	0.17	***	-0.01	0.15		0.18	0.09	**
3	0.58%	0.24%	**	-0.03	0.07		-0.27	0.13	**	-0.27	0.11	**	0.09	0.07	
4	0.72%	0.19%	***	-0.17	0.05	***	-0.25	0.1	**	-0.27	0.09	***	0.11	0.05	**
5	0.40%	0.17%	**	-0.07	0.05	†	-0.16	0.09	*	-0.22	0.08	***	0.15	0.05	***
6	0.18%	0.17%		-0.16	0.05	***	-0.14	0.09	†	-0.21	0.08	***	0.11	0.05	**
7	0.43%	0.16%	***	-0.08	0.04	*	-0.26	0.09	***	-0.19	0.08	***	0.08	0.04	*
8	-0.16%	0.23%		0	0.07		-0.1	0.13		-0.25	0.11	**	0.13	0.07	*
9	-0.27%	0.27%		0.07	0.08		-0.11	0.15		-0.27	0.13	**	0.26	0.08	***
10	-0.01%	0.37%		0.05	0.11		0.01	0.2		-0.07	0.18		0.05	0.11	

Tab. 1.9: **Long-short sentiment score equally weighted portfolio across momentum deciles:** This table shows the coefficient, standard error and p value from the regressing monthly long short sentiment portfolio return on Fama-French 3 factors and UMD. The portfolios are formed by sorting on past sentiment score within each momentum decile. Decile 1 is negative momentum decile. The portfolios are long deciles 9, and 10 and short deciles 1, and 2. The portfolios are held for five months, each month having five vintages and the returns are calculate by averaging over these vintages each month.
***: 1%, **: 5%, *: 10%,†: 15%.

SentGrp	Return		FormSent		Size		B/M		Turnover		Newscount	
1	-5.00%	0.80%	-0.61	0.03	10183	730	0.56	0.02	0.34	0.01	1.5	0.1
2	-3.50%	0.90%	-0.49	0.05	11183	884	0.52	0.02	0.32	0.01	1.4	0.1
3	-1.70%	0.80%	-0.5	0.04	12643	1039	0.52	0.02	0.3	0.01	1.5	0.0
4	-0.80%	0.90%	-0.27	0.04	15130	1036	0.53	0.02	0.32	0.01	1.6	0.1
5	1.20%	0.90%	-0.03	0.03	18723	1711	0.53	0.02	0.28	0.01	1.6	0.0
6	2.10%	1.00%	0.19	0.04	16942	1477	0.55	0.03	0.28	0.02	1.5	0.1
7	3.60%	0.90%	0.39	0.03	18253	1800	0.55	0.02	0.27	0.01	1.4	0.1
8	4.80%	0.90%	0.54	0.03	16582	1301	0.50	0.02	0.27	0.01	1.5	0.1
9	5.20%	0.90%	0.66	0.02	16397	1516	0.49	0.01	0.27	0.01	1.5	0.1
10	5.00%	0.80%	0.71	0.02	15626	1483	0.50	0.02	0.28	0.01	1.5	0.1

Tab. 1.10: **Summary statistics across analyst-sentiment deciles:** This table shows the average return, formation sentiment,formation market cap(in million USD), book to market, turnover, and newscount by formation sentiment deciles. The sentiment information is derived from news that attributes analysts. The portfolios are formed by sorting on the average of sentiment (Probability of being positive - Probability of being negative). The average of formation variable is taken over three months. The table also show standard error for each of these variables.

Newscount Group	ret		vwret		Formnewscount		size		Turnover		Newscount		Sentpos		Sentneg	
1	0.58%	0.56%	0.38%	0.54%	-0.62	0.00	4420	100	0.18	0.01	2.11	0.15	0.32	0.01	0.24	0.01
2	0.69%	0.62%	0.42%	0.53%	-0.48	0.00	2558	60	0.16	0.00	1.47	0.06	0.31	0.01	0.25	0.01
3	0.76%	0.66%	0.50%	0.53%	-0.39	0.00	2281	57	0.17	0.00	1.54	0.08	0.30	0.01	0.26	0.01
4	1.18%	0.74%	0.43%	0.54%	-0.32	0.00	1756	53	0.16	0.00	1.26	0.06	0.30	0.01	0.26	0.01
5	0.94%	0.73%	0.40%	0.50%	-0.24	0.00	1910	70	0.17	0.00	1.41	0.07	0.30	0.01	0.26	0.01
6	0.80%	0.74%	0.40%	0.51%	-0.15	0.00	2048	81	0.17	0.01	1.56	0.07	0.30	0.01	0.27	0.01
7	0.74%	0.68%	0.54%	0.49%	-0.03	0.00	2649	93	0.20	0.00	2.17	0.09	0.30	0.01	0.27	0.01
8	0.69%	0.69%	0.30%	0.52%	0.15	0.00	2914	85	0.21	0.00	2.59	0.12	0.29	0.01	0.28	0.01
9	0.70%	0.74%	0.61%	0.47%	0.45	0.01	3791	64	0.23	0.00	3.49	0.20	0.28	0.01	0.28	0.01
10	0.59%	0.82%	0.29%	0.45%	1.59	0.02	14964	304	0.28	0.01	11.60	0.65	0.26	0.01	0.30	0.01

Tab. 1.11: **Summary statistics across formation size-adjusted newscount deciles:** This table shows the average equally weighted return, value weighted return,formation standardized newscount, market capitalization (in million), turnover, newscount, standardized newscount, average probability of being positive and the average probability of being negative by standardized newscount groups.The portfolios are formed by sorting on the average of standardized newscount and held for five months. The average is taken over three months. The table also shows standard errors for each of these variables.

Newscount Group	Intercept			Mktrf			HML			SMB			UMD		
1	0.05%	0.10%		1.01	0.03		-0.04	0.05		0.44	0.05	***	-0.04	0.03	
2	-0.01%	0.11%		0.99	0.03		0.15	0.06	**	0.78	0.05	*	-0.06	0.03	*
3	0.00%	0.15%		1.06	0.04	†	0.1	0.08		0.82	0.07	***	-0.07	0.04	**
4	0.49%	0.23%	***	1.13	0.07	**	0.07	0.13		0.94	0.11	***	-0.09	0.07	
5	0.38%	0.23%	**	1.1	0.07	†	-0.12	0.13		0.95	0.11	***	-0.12	0.07	*
6	0.44%	0.22%	**	1.13	0.06	**	-0.27	0.12	**	0.85	0.1	***	-0.25	0.06	***
7	0.02%	0.12%		1.11	0.03	***	0.13	0.07	**	0.84	0.06	***	-0.05	0.03	†
8	-0.09%	0.13%		1.11	0.04	***	0.12	0.07	**	0.9	0.06	***	-0.1	0.04	***
9	-0.03%	0.14%		1.16	0.04	***	0.07	0.08		0.92	0.07	***	-0.12	0.04	***
10	-0.09%	0.17%		1.26	0.05	***	0.17	0.09	**	0.98	0.08	***	-0.29	0.05	***

Tab. 1.12: **Long only portfolios across size-adjusted newscount deciles:** This table shows the coefficient, standard error and p-values from regressing monthly portfolio returns in excess of risk free rate on Fama-French three factors and momentum. The portfolios are formed by sorting on past three months standardized newscount into deciles. The portfolios are held for five months, with each month having five vintages. The returns are calculate by averaging over these vintages each month. The p-value for excess market return is reported by comparing against the default value of 1.
***: 1%, **: 5%, *: 10%,†: 15%.

Momentum Group	Intercept			Mktrf		HML			SMB			UMD		
1	0.53%	0.31%	*	-0.10	0.09	-0.06	0.17		-0.05	0.15		0.05	0.09	
2	0.17%	0.15%		0.00	0.04	-0.07	0.08		-0.03	0.07		0.07	0.04	*
3	0.19%	0.18%		0.05	0.05	-0.09	0.10		-0.04	0.08		0.07	0.05	†
4	0.00%	0.15%		0.05	0.04	0.00	0.08		0.01	0.07		0.10	0.04	***
5	-0.04%	0.12%		-0.01	0.03	-0.01	0.06		0.05	0.06		0.08	0.03	***
6	0.28%	0.12%	**	0.01	0.04	0.00	0.07		0.05	0.06		0.01	0.03	
7	0.00%	0.10%		0.03	0.03	-0.07	0.06		0.09	0.05	*	-0.01	0.03	
8	0.29%	0.13%	**	0.01	0.04	-0.08	0.07		0.11	0.06	*	0.01	0.04	
9	0.37%	0.15%	**	0.01	0.04	-0.02	0.08		-0.01	0.07		-0.08	0.04	*
10	0.80%	0.20%	***	-0.05	0.06	-0.17	0.11	†	0.11	0.10		-0.13	0.06	**

Tab. 1.13: **Long-short size-adjusted newscount equally-weighted portfolio across momentum deciles:** This table shows the coefficients, standard errors and p-values from regressing monthly long short portfolio returns on Fama-French three factors and UMD. The portfolios are formed by sorting on past 3 month average sentiment within each momentum decile. The portfolios are long deciles 4,5, and 6 and short deciles 1,2,9, and 10 within each momentum decile. The portfolios are held for five months, with each month having five vintages. The returns are calculate by averaging over these vintages each month. ***: 1%, **: 5%, *: 10%,†: 15%.

Illiquidity Group	Intercept			Mktrf			HML			SMB			UMD		
1	0.24%	0.20%		0.13	0.06	**	-0.39	0.11		-0.14	0.10		0.29	0.06	**
2	0.17%	0.21%		0.26	0.06	***	-0.27	0.11	**	-0.29	0.10	†	0.32	0.06	***
3	0.31%	0.18%	*	0.01	0.05		-0.12	0.10	†	-0.24	0.09		0.41	0.05	***
4	0.29%	0.19%	†	-0.02	0.05		-0.12	0.10		-0.27	0.09		0.28	0.05	***
5	-0.04%	0.18%		-0.06	0.05		-0.01	0.10		-0.23	0.09		0.10	0.05	***
6	0.35%	0.24%	†	-0.21	0.07	***	-0.20	0.13	**	-0.10	0.11		0.11	0.07	
7	0.69%	0.42%	*	-0.20	0.12	*	0.00	0.23		0.09	0.20		0.19	0.12	
8	1.01%	0.31%	***	-0.33	0.09	***	-0.17	0.17	**	0.07	0.15		0.10	0.09	***
9	0.64%	0.41%	†	-0.35	0.11	***	-0.26	0.22	*	0.09	0.20		0.26	0.12	
10	-0.50%	0.83%		-0.06	0.23		0.31	0.45	**	-0.09	0.40	*	0.60	0.23	†

Tab. 1.14: **Long-short sentiment-score equally-weighted portfolio across illiquidity deciles:** This table shows the coefficient, standard error and p-values from regressing monthly long short sentiment portfolio returns on Fama-French 3 factors and UMD. The portfolios are formed by sorting on past 3 month average sentiment witin in each Amihud illiquidity decile. Group 1 is most liquid stocks. The portfolios are long deciles 9, and 10 and short deciles 1, and 2 within each illiquidity decile. The portfolios are held for five months, with each month having five vintages. The returns are calculate by averaging over these vintages each month.
***: 1%, **: 5%, *: 10%,†: 15%.

Illiquidity Group	Intercept			Mktrf			HML			SMB			UMD		
1	0.03%	0.12%		-0.07	0.04	**	0.08	0.07		0.06	0.06		0.08	0.04	**
2	-0.30%	0.13%	**	0.10	0.04	***	-0.17	0.07	**	-0.09	0.06	†	0.16	0.04	***
3	0.01%	0.12%		-0.05	0.03		-0.09	0.06	†	0.00	0.06		0.14	0.03	***
4	0.04%	0.13%		-0.09	0.04	**	-0.03	0.07		-0.02	0.06		0.13	0.04	***
5	-0.25%	0.14%	*	-0.07	0.04	*	-0.03	0.07		-0.09	0.07		0.19	0.04	***
6	0.21%	0.14%	†	-0.04	0.04		-0.18	0.08	**	-0.09	0.07		0.05	0.04	
7	0.25%	0.20%		-0.08	0.06		-0.15	0.11		-0.01	0.09		0.05	0.06	
8	0.52%	0.22%	***	0.05	0.06		-0.30	0.12	**	0.08	0.11		-0.16	0.06	***
9	0.48%	0.34%		0.10	0.10		-0.33	0.18	*	-0.09	0.16		0.01	0.10	
10	0.69%	0.24%	***	0.10	0.07	†	-0.25	0.13	**	0.20	0.11	*	-0.11	0.07	†

Tab. 1.15: **Long short size-adjusted newscount equally-weighted portfolio across illiquidity deciles:** This table shows the coefficient, standard error and p-values from regressing monthly long short newscount portfolio returns on Fama-French 3 factors and UMD. The portfolios are formed by sorting on the basis of size adjusted newscount witin each Amihud illiquidity decile. The portfolios are long deciles 4, 5,and 6 and short deciles 1,2,9, and 10. The portfolios are held for five months, with each month having five vintages. The returns are calculate by averaging over these vintages each month.
***: 1%, **: 5%, *: 10%,†: 15%.

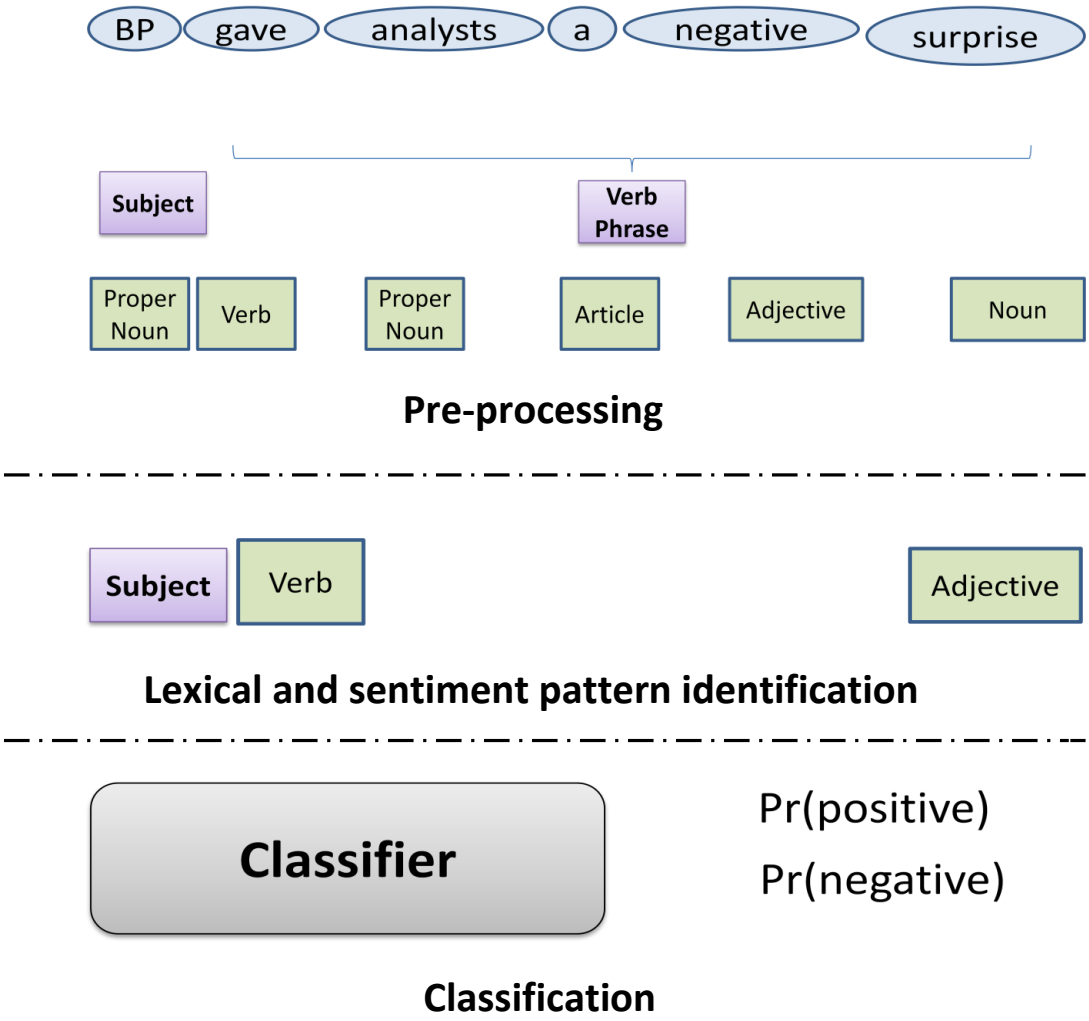

Fig. 1.2: Schematic representation of the sentiment engine:

This figure shows how an example sentence passes through the three different subprocesses of the Reuters sentiment engine.

TIMESTAMP	RIC	SentPos	SentNeg	Relevance	Item_Type	Lcnt	TopicCd
2006-01-03 12:45:10	PPC.N	0.06	0.81	1.00	Article	2	FOD US RES
2006-01-03 12:45:35	F.N	0.55	0.01	1.00	Alert	2	AUT RCH RTRS
2006-01-03 12:45:50	LCC.N	0.11	0.72	0.41	Article	1	AIR RCH RTRS

Fig. 1.3: Example of select fields of the sentiment engine:

The figure shows select fields of the raw database. This example shows newsitems for Pilgrim's Pride, Ford and US Airways respectively. The news item types are article, alert and article respectively. The news items have RES (results), RCH (research) and RCH (research) respectively.

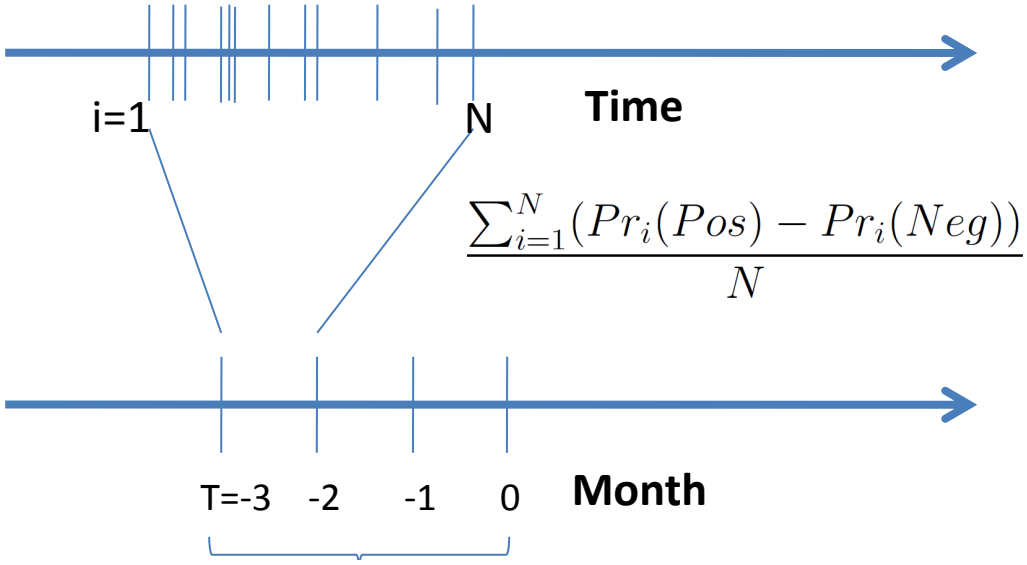

Fig. 1.4: Aggregation of news to obtain historical sentiment:

Each month a monthly sentiment score is obtained by taking the average over all the articles. From the monthly scores, a measure of long term sentiment is obtained by taking three-month moving average of monthly sentiment scores.

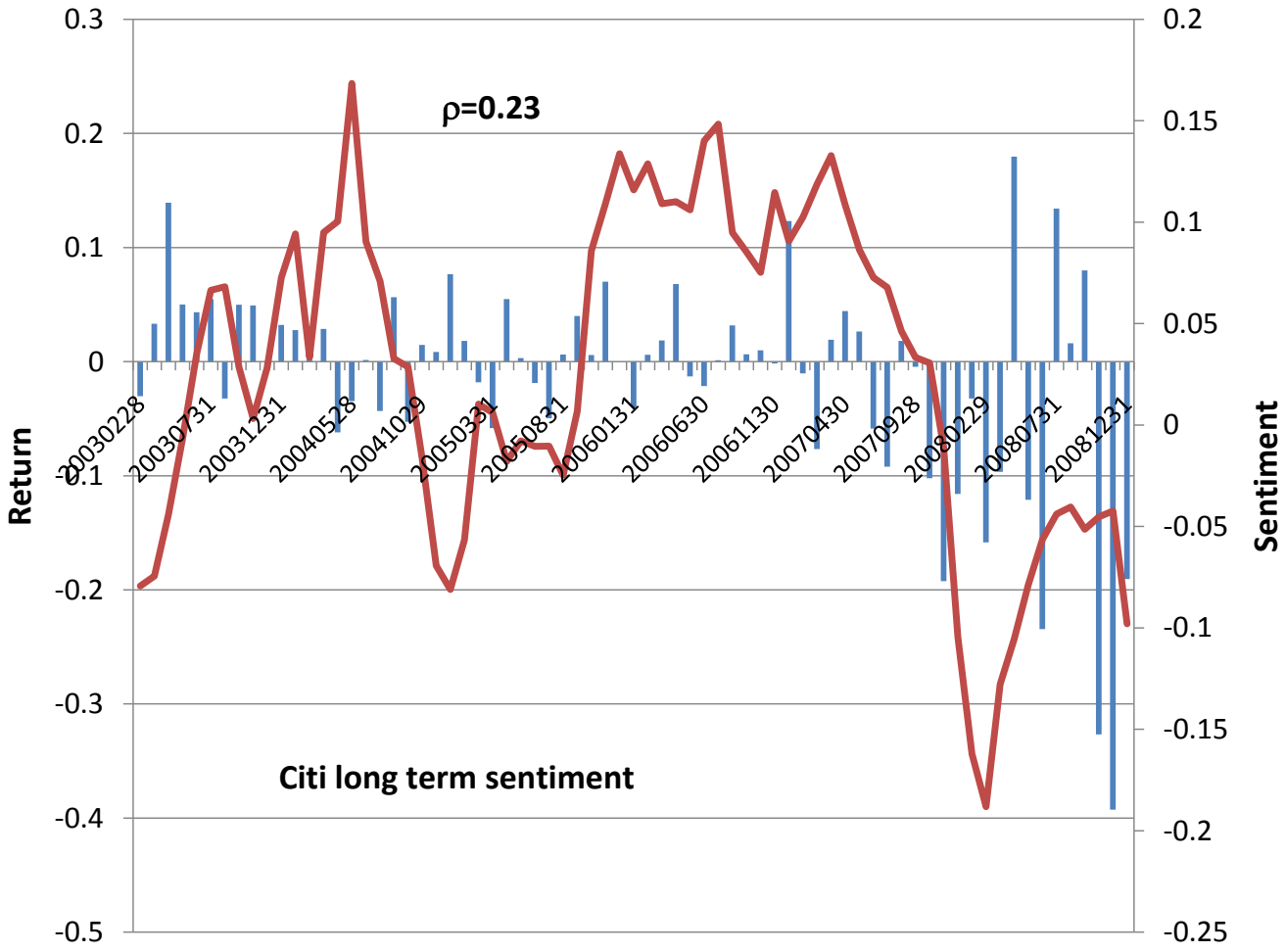

Fig. 1.5: Example of long term sentiment score:

The solid line on the chart shows how Citi (Ticker: C) long term sentiment changed during the sample period. Citi had two periods of negative news -(1) due to uncertainty in the CEO succession, (2) during the financial crisis. The long term sentiment captures both. The bar shows how the monthly return for the stock varied during the same time period. The correlation between long term sentiment and monthly return is 0.23, as reported on the chart.

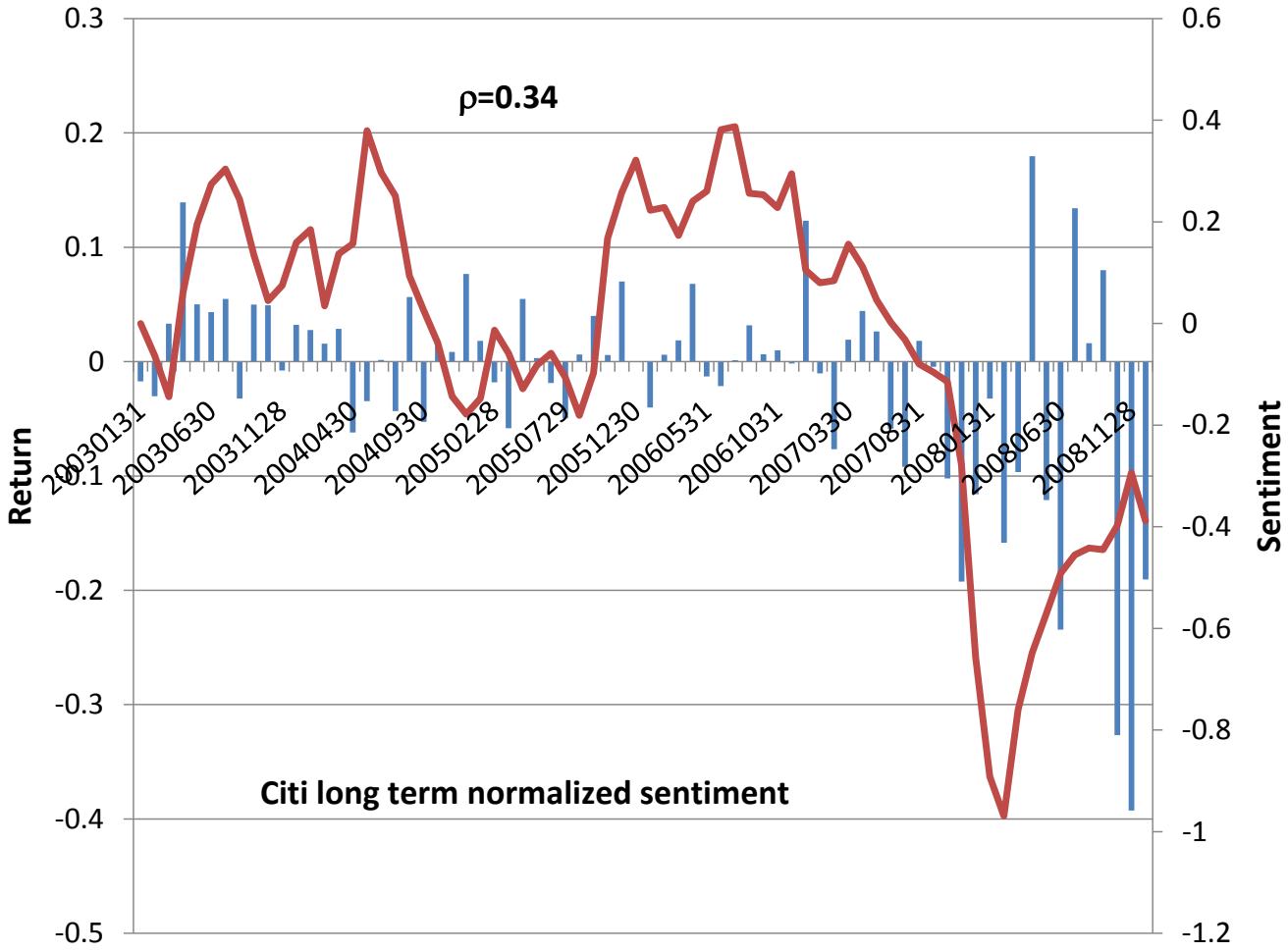

Fig. 1.6: Example of normalized long term sentiment score:

The solid line on the chart shows how Citi (Ticker: C) normalized long term sentiment changed during the sample period. The normalization is carried out by first subtracting cross sectional long term sentiment and then dividing by the standard deviation of long term sentiment. Citi had two periods of negative news- (1) due to uncertainty in the CEO succession, (2) during the financial crisis. The long term sentiment captures both. The chart also shows that the long term sentiment surrounding CEO succession was less negative than during the financial crisis. The bars show how the monthly return for the stock varied during the same time period. The correlation between long term sentiment and monthly return is 0.34, as reported on the chart.

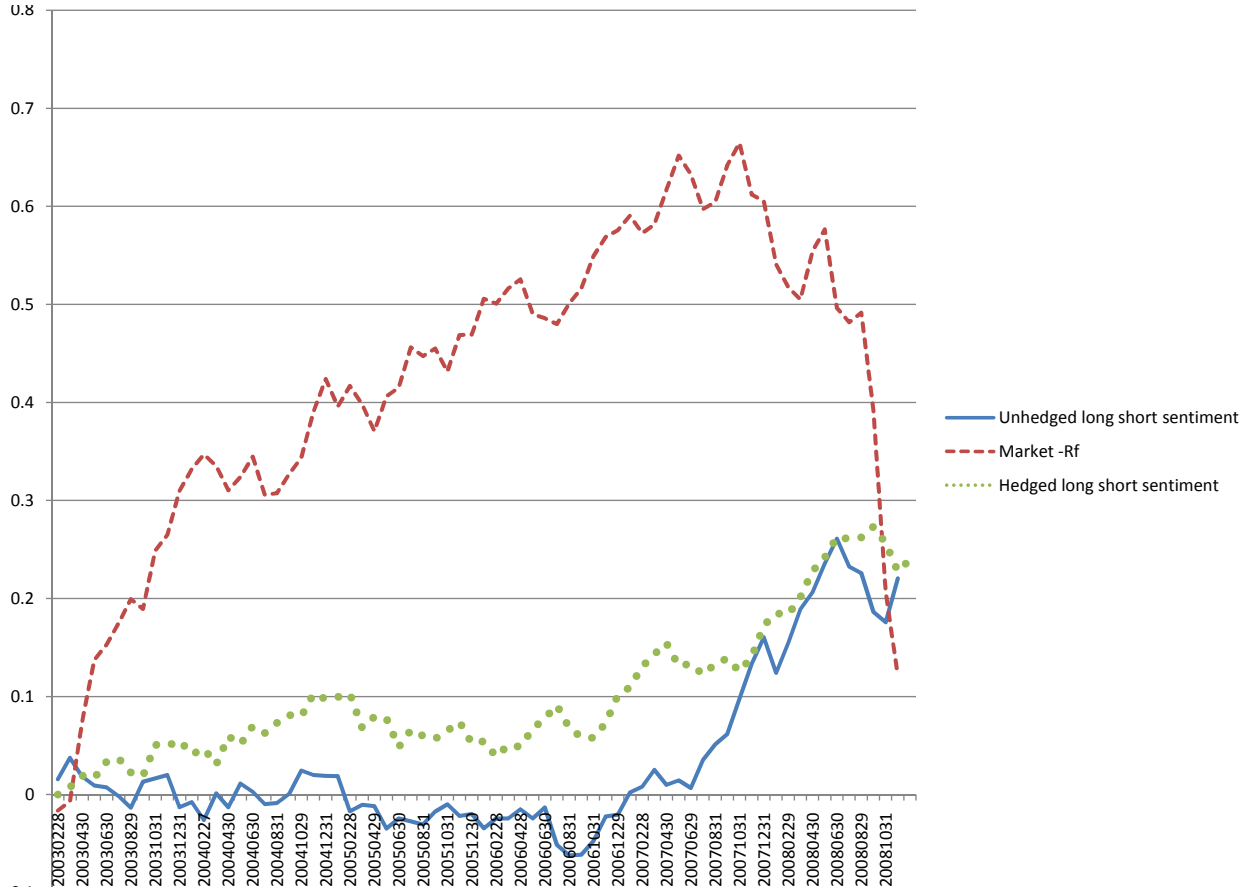

Fig. 1.7: Performance of long short sentiment trading strategy over time

The solid line in the figure shows the performance of the long-short sentiment portfolio during the sample period. The portfolio begins with zero dollars and ends with 22 cents. The green broken line in the figure shows hedged long-short portfolio, while the red broken line shows the excess return on the market portfolio. The hedged portfolio return is obtained by taking out the exposure to the Fama-French three factors and UMD.

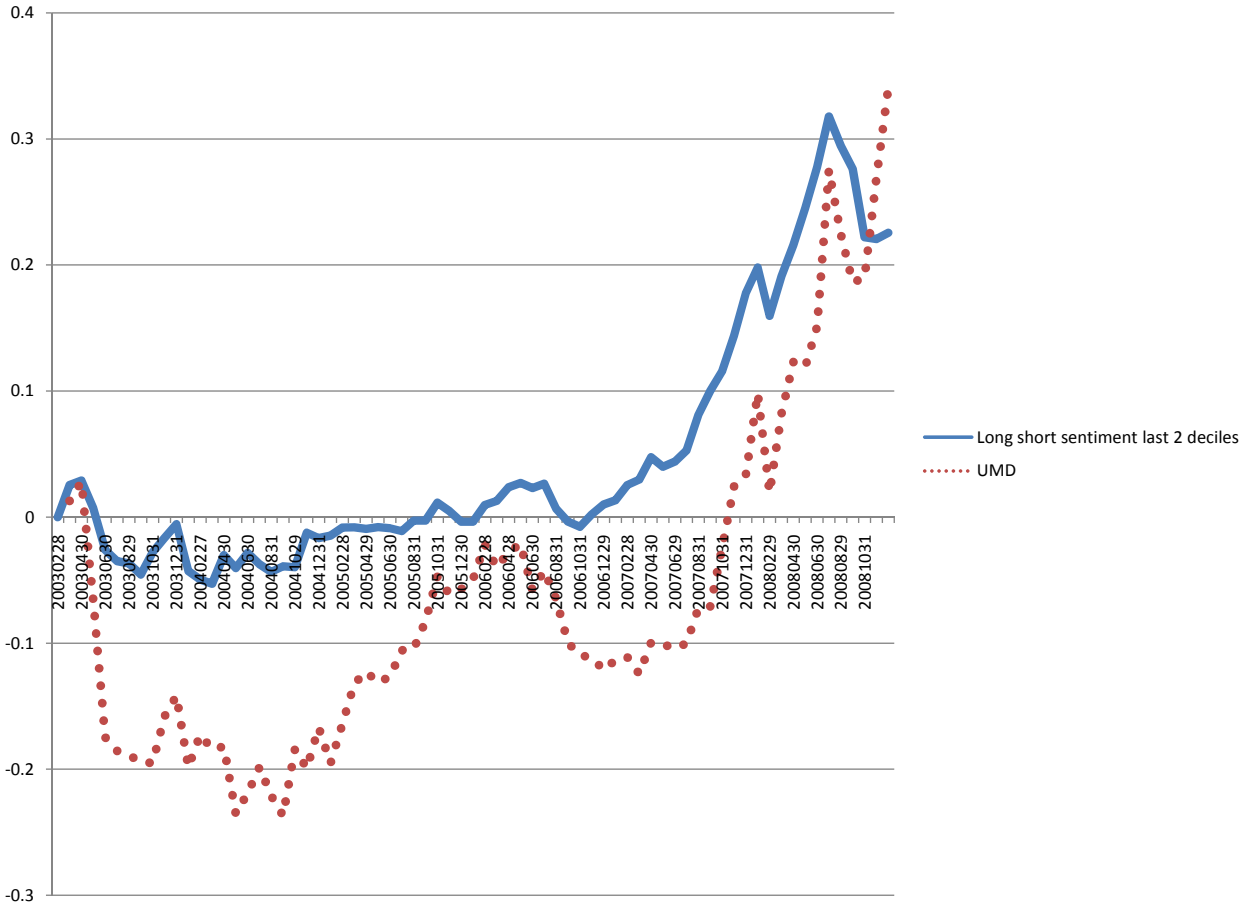

Fig. 1.8: Comparison of long short sentiment trading strategy with UMD over time

The solid line shows the performance of the long-short sentiment portfolio when applied to the largest quintile of firms in the sample. The shaded line is the performance of the UMD factor during the sample period. The figure shows how the two portfolios are highly correlated during the sample period.

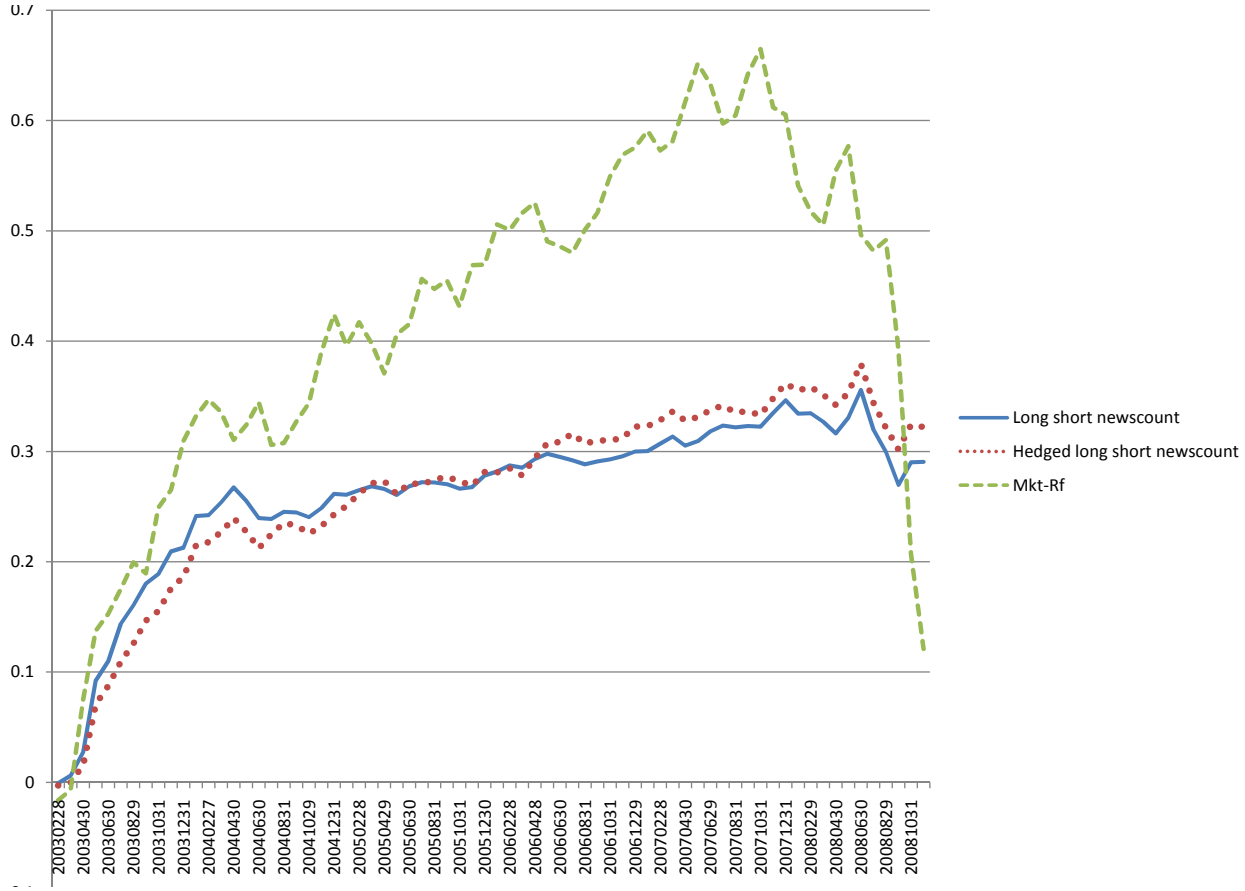

Fig. 1.9: Performance of the size-adjusted newscount trading strategy over time

The solid line in this figure shows the performance of the long-short size-adjusted newscount portfolio during the sample period. The portfolio begins with zero dollars and finishes at 29 cents. The green broken line shows the hedged portfolio. The shaded red line shows the excess market return during the sample period.

2. NEWS VOLUME AND TRADING ACTIVITY

2.1 Introduction

[1] It is not unusual to see almost one news item per day for IBM on a trading terminal and hardly any news for Xilinx. The amount of time that goes by before one observes some information about a stock varies from stock to stock. This paper shows that the cross-sectional variation in news articles across stocks is related to trading activity (expressed as product of dollar volume and volatility) in a very precise manner. Based on the intuition that the "time clock" for different stocks ticks at different rates, proportional to the rate at which "bets" are placed, the model of Kyle and Obizhaeva (2009) suggests the hypothesis that a one unit increase in trading activity translates into a two-thirds of one percent increase in the arrival rate of news articles. We find support for this relationship using a dataset of Thomson Reuters news articles. The relationship is robust to various models of count data. The relationship is also robust to excluding various type of firm specific news. By showing this relationship, the paper highlights the possibility that the unit of time in market micrcostructure is trading time instead of calendar time.

In models of market microstructure, the role of time is understudied. O'Hara

[1] This chapter is result of joint work undertaken with Albert S. Kyle, and Anna Obizhaeva. We thank Tugkan Tuzun for providing data on trading activity.

(1995) states that "the importance of time is ultimately an empirical question . . . ". We start from the theoretical model of Kyle and Obizhaeva (2009) in which the market microstructure "fundamentals" are similar across all stocks, but the "time clock" for more actively traded stocks clicks faster than the time clock for less active actively traded stocks. This difference in speed of time clock leads to empirically observed differences. Kyle and Obizhaeva (2009) call this as Time-Clock Irrelevance. Time-Clock Irrelevance argues that speeding up the trading game does not change the structure of the game. As a result of a faster time clock, "traders will make economically equivalent trades at economically equivalent prices, but the trades and price changes will unfold proportionately faster . . . .Intuitively, time clock irrelevance refers to the idea that speeding up or slowing down a trading game does not have an effect on the fundamentals of the game." In this paper, we examine the hypothesis that the relationship between trading activity and the information flow in the market, as measured by the frequency of news articles, corresponds to that predicted by the time clock irrelevance.

Following, Kyle and Obizhaeva (2009), traders are assumed to place "bets" — defined as statistically independent decision to trade specific quantities — at a rate proportional to which the time clock ticks. "Trading activity" is defined as the product of dollar volume and percentage returns standard deviation. Incorporating standard deviation into the definition of trading activity makes trading activity a measure of risk transferred. Trading activity is also unaffected by Modigliani Miller irrelevant stocks splits and leverage changes. Consider changing a stock's time clock so that one hour is changed to H hours. If $H > 1$, the time clock has been slowed;

if H $<$ 1, the time clock has been sped up. Speeding up the time clock affects trading activity in two ways. First, speeding up the time clock (H<1) increases the number of bets per day and therefore dollar volume increases proportionately with $\frac{1}{H}$. This is called the "volume effect". Second, speeding up the time clock increases returns variance proportionately with 1/H, so volatility (the square root of variance) increases proportionately with $\frac{1}{H^{1/2}}$; as a result, speeding up the clock (H<1) makes each bet riskier. This is called the "volatility effect". Since trading activity is defined as product of dollar volume and volatility. Combining the volume effect and the volatility effect increases trading activity by the factor $\frac{1}{H^{\frac{3}{2}}}$. What does speeding up the clock do to the amount of information that arrives in an hour? We make two assumptions.

- Information arrives at a rate proportional to the rate at which the time clock ticks.
- News articles arrive at a rate proportional to the rate at which information arrives.

Under these assumptions, time clock irrelevance implies that the arrival rate of news articles increases by a factor $\frac{1}{H}$. If μ^* denotes the arrival rate of news articles when H=1, it is related to the sped up arrival rate μ by the following equation:

$$\mu = \mu^* \times \frac{1}{H} \tag{2.1}$$

Similarly, sped up trading activity W is related to the original rate of trading activity

W^* as follows.

$$W = W^* \times \frac{1}{H^{\frac{3}{2}}} \tag{2.2}$$

Combining (2.1) and (2.2) to eliminate H, we obtain the following relationship.

$$\frac{\mu}{\mu^*} = \Big(\frac{W}{W^*}\Big)^{2/3} \tag{2.3}$$

The above relationship may also be expressed as,

$$\mu = \frac{\mu^*}{W*^{2/3}} \times W^{2/3} \tag{2.4}$$

Equation (2.4) identifies the relationship we test in this paper. The equation states that the elasticity of the arrival rate of news articles with respect to trading activity W is two-thirds. Thus, a one percent increase in trading activity will be accompanied by a two-thirds of one percent increase in the arrival rate of news articles.

Let us consider a numerical example. If $H = \frac{1}{2}$, the same amount of information that used to arrive in an hour now comes in half an hour. Dollar volume goes up by a factor of 2, since each day investors will trade twice as many shares. In our numerical example, the variance doubles or equivalently, the standard deviation increases by $\sqrt{2}$. Trading activity, which is the product of dollar volume and volatility,goes up by a factor of 2.8 ($= 2^{\frac{3}{2}}$). Kyle and Obizhaeva (2009) argue that Modigliani-Miller irrelevance makes the particular dollar volume factor of 2 and the standard deviation factor of $\sqrt{2}$ irrelevant; the economically important number is their product $2^{\frac{3}{2}}$.

What is the intuition for the length of trading day? The length of the trading day is proportional to the rate at which stock trading fundamentals unfold. Importantly, this includes the amount of information that arrives in a unit of calendar time. Trading days are defined so that in each trading day of length H (which varies across stocks), the rate of information arrival is the same across all the stocks. In such a set up a trading hour is not the same physical time for all stocks; instead, it is a long physical time if the trading day is a physical minute and very short physical time if the trading day is physical eight hours. A model for duration between information arrival and a model for the number of information arrival rates are dual because a Poisson arrival rate is reciprocal of the number of arrivals in one unit of time.

The rate at which information arrives is hard to quantify. In this paper, we use as a proxy for information arrival rate the rate at which public news arrives at trading desks. It is not easy to observe the true arrival rate of information, since it is an unobserved latent variable. We infer the sensitivity of the arrival rate to trading activity by modeling the arrival rate of news articles as a function of trading activity. The time clock irrelevance model of Kyle and Obizhaeva (2009) suggests the following testable hypothesis: for each one percent increase in trading activity, there is a two-thirds of one percent increase in the arrival rate of news article. Thus, we test the hypothesis that the arrival rate of news article is proportional to $W^{\frac{2}{3}}$, where W denotes trading activity.

Both the time clock irrelevance and the hypothesis that information arrives at the rate the time clock ticks, relate to the proposal to model returns as a subordi-

nated Brownian motion in local time by Mandelbrot and Taylor (1967) and Clark (1973). Local time corresponds to a measure of trading volume or transaction time. The concept of a transaction is similar to the concept of a bet.

Our specific proxy for information is the number of news articles in a Thomson Reuters database of firm-specific news articles. Thomson Reuters news articles appear on traders' screen in real time. As our proxy for the information arrival rate, we count the number of news items about each firm for each month. Sometimes, the same news article is informative on multiple dimensions. For example, the same news item may announce both earnings and an impending merger. We count such a news article about two issues (earnings and merger) either as one or as two units of news. Sometimes, the same news item is repeated with some added information. We count this as two news items. The task of counting the number of news articles is made simpler by Thomson Reuters' topic codes provided in their database of news articles. Topic code indicates the topic of the news article, e.g. earnings, earnings forecasts, mergers, rating changes, corporate filings. For each firm, we investigate both counting the article as one news item and counting the article as the number of topic codes mentioned in the article.

This paper relates to the studies by Mitchell and Mulherin (1994) and Berry and Howe (1994), who show that arrival of news affects market activity and conclude that there is a weak cross-sectional relationship between news and market activity. Our paper shows a strong cross-sectional relationship. One possible reason for the difference between our paper and others is that the we combine volume or volatility into one measure of trading activity. More recently, Tetlock (2009) uses

news data from the Dow Jones archive to investigate the idea that public information reduces information asymmetry. Our paper assumes that the amount of public news is proportional to the amount of total information available. At first glance, it may appear unlikely that the amount of public news is proportional to the amount of private information, but there are good reasons to believe so. First, private information may arise due to the manner in which public information is processed. Second, news reporters may write articles about the same firms for which traders are starting to acquire private information.

This paper provides some explanation for factors that affect the cross-sectional variation in trading activity documented by Chordia, Huh, and Subrahmanyam (2007). Our paper shows that cross sectional variation in trading activity is related to cross sectional variation in information. Chordia, Huh, and Subrahmanyam (2007) point out that factors that influence visibility of the stock, or indicate dispersion of opinion, influence cross sectional differences in trading activity.

This paper also relates to Dufour and Engle (2000), who show empirically that as the time duration between transactions decreases, the price impact of trades increases. This paper suggests that a decrease in time duration between trades corresponds to faster information production, hence increased trading activity and price impact. This paper also suggests that there is a role played by the cross-sectional difference in arrival rate of firm specific news, hitherto unexplored.

This paper aggregates different types of news arriving on a trader's terminal. Additions and deletions from stock-indices, new listings, delistings and suspensions, mergers and acquisitions, earnings announcements, and results forecasts constitute

over 50% of firm specific news items that a professional trader receives on a terminal. Another 10% of news items is a surprising news (or breaking news). Analysis of corporations by reporters and debt market news make another 13% together. News about analyst upgrades and downgrades make 4.8% of our sample. This is in sharp contrast to the amount of emphasis some market participants put on analyst revisions.

The log-linear relationship between expected number of news articles and trading activity can be thought of a power law. In this sense, the paper relates to literature on power laws in finance. Relationships between many variables have been found to follow power laws, i.e., relationships in which one variable is proportional to a power of another variable. Gabaix, Gopikrishnan, Plerou, and Stanley (2003) discuss power laws in financial markets. Usually the existence of a power law is a challenge to explain. In this paper, cross sectional variation in the speed of the time clock provides the link between trading activity and the amount of information for a stock.

Section 2.2 describes the data. Section 2.3 describes the estimation methodology. Section 2.4 discusses the results. Section 2.5 Concludes.

2.2 Data

The news data is obtained from the Thomson Reuters NewsScope dataset from January 2003 to December 2008. For news articles mentioning specific firms, Thomson Reuters provides both a ticker symbol and one or more firm-specific topic

codes. Topic codes are journalist-and-computer generated tags which indicate a particular kind of event. For example, all news items which relate to an earning announcement are tagged as 'RES'. All news items provided by Thomson Reuters are tagged as 'RTRS'. The tag of 'RES' is created by a journalist, while 'RTRS' is automatically generated by a computer. The item also contains information on whether a news item is a news 'article' or an 'alert'. 'Alerts' are one-line headlines followed later by an 'article'. We only consider articles in our sample.

A news item can have information on multiple dimensions of the firm. For example, the same news item could be an earnings announcement, a management forecast, and a merger announcement. If such a multidimensional news item is counted as one, it potentially underweights the amount of information available in the news relative to a news item with one topic code. One way to get around potential underweighting of information content is to count the above-mentioned news item as three items: first as a news item on earnings announcement, second as a news item on an earnings forecast and finally as a news item on a merger announcement. Thomson-Reuters-provided tags indicate the dimensions that the news is informative about. We start with a comprehensive list of all tags used by Thomson Reuters. From these tags, we cull out tags that relate to firm specific news. We omit tags that contain information about an industry or the firm but are not matched to a firm-specific ticker symbol.

2.2.1 What constitutes news

To create a comprehensive database of firm specific news items, a list of tags was created and all the news articles containing those tags were included. The news database was matched with a database of common stocks that are traded on NYSE, Amex and NASDAQ. A listing and brief description of these tags as well as the proportion of news articles for each tag in the matched data is provided in table 2.1.

Across all the topic codes, we obtain 3.4 million topic code mentions. Almost 1.4 million news stories constitute the database. Approximately 80% of news tags are the following.

- Almost 15% of the news articles contain the topic code 'STX', which indicates additions and deletions from stock indices, all new listings, delistings and suspensions.[2]

- Almost 14% of news articles contain topic code 'RES', which indicates annual and quarterly earnings.

- Almost 12% of news items contain the topic code 'MRG', which indicates mergers and acquisitions.

- Almost 9.7% of news items contain the topic code 'RESF', which indicates results forecast.

- Almost 9.4% of news items contain the topic code 'NEWS', which indicates news *that is likely to lead television or radio bulletins or to make the front*

[2] 'STX' is not used for broker research; news about individual companies unless the news is about addition, deletion, new listing or suspension; and hot stocks.

pages of major international newspapers and web sites.

- Almost 6.7% of news items contain the topic code 'CORA', which indicates analysis of a company by a journalist.

- Almost 6.5% of news items contain the topic code 'DBT', which indicates news related to debt markets.

- Almost 4.8% of news items contain the topic code 'RCH', which indicates news about broker research.

- Almost 4.9% of news items contain the topic code 'HOT', which indicates news about stocks that are on move.

We aggregate the news data by summing the number of tags for each firm for each month for the years 2003-2008. We have 72 months of data. In course of the 72 months, we obtain 275,059 firm-month observations resulting from at least one match between a firm and a news article topic code. If Thomson Reuters had uniform coverage for our sample, we should have almost 1.38% ($\frac{1}{72}$)of our observations each month. In fact, the coverage increased over time, as a result of which our data is weighed more towards the later periods. Figure 2.1 shows that the coverage of Thomson Reuters increased each year during our sample period. Across the 72 months, the average number of firm each month is 3,820. We have at least 2,586 firms in our sample in each month. We count a firm as covered by Thomson Reuters from the instance we observe the first news item for the firm in our sample. For some firms, the rate of news arrival is slow enough that firms show up with fewer

than one observation per month.

We match the firm's ticker symbol with aggregated with daily returns, price, and volume data obtained from CRSP. Tables 2.2 and 2.3 provide summary statistics for the matched dataset as well as for 10 volume groups. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume. Table 2.2 presents sample statistics as well as statistics by volume group. If different topic codes are not counted multiple times, for the entire sample, the average number of news items per firm per month is 4.2. There is significant variation in newscount across volume groups. For the lowest volume group, the monthly newscount is 0.6 news items per month, while that for the highest group is 82.5 news items per month. For the over all sample, almost 58% of firms have no news in a particular month. For the highest volume group the proportion of firms without any news in a particular month is only 4.84%.

2.2.2 Count by topic code

If a news item has multiple topic codes mentioned, it can be counted multiple times, once for each topic code. For example, if a news article is tagged as having information about both an impending merger and an earnings forecast, it can be treated as two news items. Counted this way, for the entire sample, the median number of newscount per month is 0 and the average is 1. For the highest volume group the median monthly newscount is 84. If we exclude news items about div-

idends, the median newscount for the highest volume group is 74 news items per month, while for the whole sample it is 0. If we exclude the topic code 'NEWS' as well, the median for higherst volume group is 63 news items per month while for all firms it is 0 news items per month. For the overall sample, 39% of the firms have zero news items in a given month. For the highest volume group only 2.2% have no news in a given month. Median statistic are always lower than the mean. This suggests that there are some firms with extremely large numbers of news articles. This fact is also consistent with large values for the maximum number of news items in each volume category.

Figure 2.2 shows the distribution of news items per month for our sample. Since there are some firms with large numbers of news items, the data is modified as follows. News counts between 50 and 75 are recoded as 50, between 75 and 100 are recoded as 51, between 100 and 150 as 52, between 150 and 300 as 53, between 300 and 500 as 54 and more than 500 are coded as 55.

Figure 2.3 shows the distribution of news items excluding dividends per month, following the same format as figure 2.2. Two important facts stand out in figures 2.2 and 2.3: (1) Both figures have relatively large number of zeros and (2) If we do not count dividends, the number of firms with zero news items in a month is higher (48% in the first, 54% in the later).

Table 2.4 shows the distribution of news items per month up to 10 news items. Unlike table 2.3, where monthly averages are presented, the figure treats the entire dataset as a cross section. About 48% of the firms in each month have zero news items in a month. If we do not count dividend as news, almost 54% firms do not

have a news item in a month. This has significant implications for the model of news arrival rate estimated below.

2.3 Estimation

We aim to explore the relationship between the arrival rate of information and trading activity. The arrival rate of information is unobservable; we observe the number of news articles per month instead. For some firms, the arrival rate is low, as evidenced by relatively large number of firms who receive no news items in a given month.

For each stock i with trading activity W_i, we observe N_i news per month. Variable N_i is a count variable with possible zeros.

We implement three estimation approaches -(1) a log-linear model, (2) a Poisson model, and (3) a negative binomial model.

- **Log-linear model:** Since we are estimating the mean arrival rate of random count data, it makes sense to estimate models based on Poisson arrival rates as we do below. A simpler approach is to group firms based on trading activity W and instead estimate a simple log linear model which estimates how the average number of news items per stock varies as a power of trading activity. This simple approach has two statistical advantages: (1) averaging over many firms reduces statistical variation in the number of Poisson arrivals, (2) there is no need to take logs when firms have zero news articles. To accomplish this, we sort all the firms by trading activity for each month. For each month, we

form 30 groups constructed so that each group has the same number of news items. We obtain the average number of news items per firm in each group and the average trading activity per firm each group. By construction, neither of these two numbers is zero. We regress the log of the average number of news items in each group on the log of the average level of trading activity in each group.

- **Poisson Model:** The Poisson model assumes that the arrival rate of news articles is a possibly non linear function of trading activity which we denote as W. This model implies that a stock i with trading volume W_i has a probability of $\tilde{N}_i$ news items in a month given by:

$$f(\tilde{N}_i|W_i) = \frac{e^{-\lambda(W_i)} \times (\lambda(W_i))_i^N}{N_i!}. \tag{2.5}$$

In the Poisson model, we estimate the arrival rate as a linear function of log(W).

$$\lambda(W) = e^{\mu+\gamma \ln W}. \tag{2.6}$$

The model of time clock invariance predicts $\gamma = \frac{2}{3}$. This Poisson model assumes that the arrival rate is a non-stochastic function of W, i.e., it assumes that all variation in arrival rates occurs within the context of the Poisson distribution. It therefore assumes that $\lambda(W) = Var(\tilde{N}_i|W_i) = E(\tilde{N}_i|W_i)$. The Poisson model assumes that stocks with the same level of trading activity have the same expected number of news articles $\lambda(W)$.

- **Negative Binomial Model:** The negative binomial model allows the Poisson arrival rate to vary randomly, even for firms having the same level of trading activity. We model the variation with a gamma distribution. Let Gamma(k,θ) denote a random variable with a gamma distribution having mean and variance of kθ, and kθ^2, respectively. Letting α denote the variance of gamma distribution, we estimate a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate is the gamma distribution:

$$\lambda(W) = e^{\mu+\gamma \ln W} \times \text{Gamma}(k, \theta). \tag{2.7}$$

In this specification, the model parameter μ identifies the same mean as the mean of the Poisson distribution. Therefore we impose the restriction $k = \frac{1}{\alpha}$ and $\theta = \alpha$. This specification restricts the mean of the gamma distribution to be 1. The estimated parameter α is the variance of the gamma distribution, $\alpha = k\theta^2 = \theta$. The negative binomial model nests the Poisson model as a special case with $\alpha = 0$. For a given mean, the negative binomial model allows the variance of the number of news articles to be greater than implied by the Poisson model. This model will therefore fit the data better if firms with similar levels of trading activity have dramatically different newscounts, too large to be explained by the Poisson model.

Kyle and Obizhaeva (2009) compare their model based on time clock invariance against two alternative models. We do the same here. The model of time clock invariance is based on the idea that traders place bets at a rate proportional to the

rate at which time clock ticks. Our tests of this model are based on the following identifying assumptions:

- Information – like bets – arrives at a rate proportional to the rate at which the time clock ticks.
- The rate of information arrival is proportional to the rate of arrival of Reuters news articles.

Together these two assumptions imply that news articles arrive at a rate proportional to the arrival rate of bets. As discussed, this implies the restriction $\gamma = \frac{2}{3}$ in equation (2.6).

In the first alternative model of Kyle and Obizhaeva (2009), the "model of invariant bet frequency", all increases in trading activity come from increases in trade size. The arrival rate of bets does not vary with trading activity. Instead traders place larger bets. Our implementation of this model assumes, therefore, the restriction of $\gamma = 0$ in equation (2.6).

In the second alternative model, "model of invariant bet size", all increases in trading activity come from increases in trade frequency and not bet size. In this model, the arrival rate of bets is proportional to trading volume. Since we assume that the arrival rate of bets is proportional to the arrival of news articles, our interpretation of the model of invariant bet size imposes the restriction $\gamma = 1$ in equation (2.6).

2.4 Model Estimation

In this section, we report the results of the estimation strategy described in previous section. We first report results when firms are sorted each month into 30 groups by trading activity. Then we test the Poisson and negative binomial model.

2.4.1 Log-Linear models with data in bins

All stocks are first sorted by W. We construct 30 groups, such that each group has the same number of news items. We then calculate the average number of news items per firm for each group and average trading activity for stocks in each group. We plot the average of log(number of news articles) against logarithm of average trading activity within each trading group. The theory proposed by Kyle and Obizhaeva (2009) implies a linear relationship with a slope of $\frac{2}{3}$.

Figure 2.4 shows the plot.

From figure 2.4, it is clear that a slope of $\frac{2}{3}$ fits the data well, but the graph has a visible "smile" indicating some convexity in the relationship we have assumed to be linear. In comparison with the fitted line, the bins with very active and very inactive stocks have "too many" news items, and the stocks in the middle have "too few" news articles.

The following model is estimated across the trading activity activity groups.

$$\log\left(\text{number of news articles}\right) = m_t d_t + \gamma_t \log(W_t) + \epsilon_{it} \tag{2.8}$$

In this specification, d_t represents a dummy variable for each month, and m_t is the

coefficient on the dummy variable.

The results of the model estimation are presented in table 2.5. The table presents four different sets of results, divided into two sets of two regressions each. In the first set of regressions, each news item is counted only once; in the second set of regressions, each news items is counted multiple times, once for each topic code. Within each set of regressions, two sub-sets of results are presented, based on different ways of including firms which Reuters does not cover. In the first set, only firms which Reuters covers are included in the data. In the second set, firms which Reuters does not cover, but would have been included if Reuters covered them, are included. By construction, the added firms are firms which never have a news article in any month. Since these firms tend to be small firms, inclusion of them tends to decrease the average number of news articles for the smaller firms.

To construct the data used in the regression, for each of the 72 months in the years 2003-2008, the firms were sorted by number of news items into 30 bins, where each bin was constructed to contain approximately the same number of news items. By construction, the bins for stocks with few news articles have large numbers of stocks. For each bin and for each month, the average number of news articles and the average level of trading activity was calculated. The estimation is an OLS regression of the log of the average number of news articles per firm on the log of average trading activity per firm, plus dummy variables for each month. Since there are 30 observations on each of 72 months of data, the regression involves 2160 observations. Including the 72 dummy variables representing fixed effects and the coefficient γ, there are 73 degrees of freedom in the regression.

As shown in table 2.5, in the four versions of the regression, all four of the regression coefficients are economically close to the predicted value of 2/3, and economically very far away from the alternative model predictions of $\gamma = 0$ and $\gamma = 1$. In the two regressions where news items are counted only once each, the parameter estimate is $\gamma = 0.65$ when only firms coverage by Thomson Reuters are included and $\gamma = 0.68$ when firms with no coverage are added. When news is counted multiple times when there are multiple topic codes, the parameter estimate increases from $\gamma = 0.65$ to $\gamma = 0.70$ when only covered firms are included and from $\gamma = 0.68$ to $\gamma = 0.74$ when non-covered firms are included. Since the standard error is 0.005, the model tends reject the hypotheses $\gamma = \frac{2}{3}$, even though the results are economically close to the predicted value of $\frac{2}{3}$. The F statistic for $\gamma = \frac{2}{3}$ range from 4 to 196. The model, where all the firms are included and each news item is counted only once, is the only one that that fails to reject $\gamma = \frac{2}{3}$.

The R^2 values range from 0.89 to 0.91. While this indicates that the linear specification explains the vast majority of the variation in average news count across bins, the "smile" in figure 2.4 suggests that much of the unexplained variance could be captured by including a quadratic term in the regression.

To investigate the hypothesis that the number of new articles referencing different types of news topics might vary differently as trading activity varies, we estimated equation (2.8) separately for news items referencing different topic codes. The results for the 14 most common topic codes are presented in table 2.6. For each topic code, representing a row in the table, a coefficient estimate for γ is presented, along with standard errors and t-stats for the three hypotheses $\gamma = \frac{2}{3}, \gamma = 0, \gamma = 1$.

Estimated coefficients across different topics range from $\gamma = 0.30$ to $\gamma = 0.96$. With the exception of the two topic code for dividends ("DIV") and news ("NEWS"), the t-statistics indicate that the hypothesis $\gamma = \frac{2}{3}$ fits the data overwhelmingly better than either alternative hypothesis $\gamma = 0$ or $\gamma = 1$. For the remaining 12 topic codes the coefficient estimates range from $\gamma = 0.54$ to $\gamma = 0.75$, much closer to $\gamma = \frac{2}{3}$ than to either $\gamma = 0$ or $\gamma = 1$. The large amount of variation in the coefficients indicates, however, that news on individual topics does not precisely fit the model of Kyle and Obizhaeva (2009). For firms with different levels of trading activity, Thomson Reuters tends to focus on different types of news.

For the topic code "DIV", the estimated coefficient is $\gamma = 0.30$, the smallest of the 14 coefficients. This fits the hypothesis $\gamma = 0$ better than $\gamma = \frac{2}{3}$. This result indicates that news articles about dividends are a larger proportion of total news articles for firms with low trading activity than for firms with high levels of trading activity.

For the topic code "NEWS", the estimated coefficient is $\gamma = 0.95$, the largest of the 14 coefficients. This fits the hypothesis $\gamma = 1$ better than $\gamma = \frac{2}{3}$. This result indicates that articles suitable for major national news outlets are a larger proportion of total news articles for firms with high trading activity than for firms with low levels of trading activity. Such news items might appear on a CNN breaking news program.

2.4.2 Count data models

The log-linear model with data in bins does not provide a statistical explanation of why, given two firms with similar levels of trading activity, one firm might have many news items in a given month and the other firm might have zero news items in the same month. The negative binomial specification of equation (2.6) allows the number of news items in a month to vary for three reasons: (1) variation in the Poisson arrival rate associated with different levels of trading activity, (2) an additional component of variation in the Poisson arrival rate associated with otherwise unmodeled features captured by the gamma distribution, (3) random variation in the actual number of Poisson events given the Poisson arrival rate determined by trading activity and the gamma distribution.

We estimate equation (2.6) in the following manner. For each of the 72 months in the period 2003-2008, the "count data" model in equation (2.6) is estimated using maximum likelihood with the procedure is SAS. The 72 parameter estimates are then averaged together and the average is reported as the parameter estimate for the model. Standard errors are calculated in three different ways: Newey-West, clustering, and monthly fixed effects.

Table 2.7 reports estimates of the constant term μ, the log-linear coefficient γ, and the gamma distribution variance parameter α for twelve different specifications of the negative binomial model in equation (2.6). There are three sets of four results each, where each of the three sets is based on four different versions of the data. The three set of results are differentiated by the following different ways of counting

news articles: (1) counting each news item once; (2) counting each each news items possibly multiple times equal to the number of topic codes; (3) counting each news items multiple times equal to the number of topic codes, but excluding the topic codes RES and DIV. For each set of results, there are two subsets of results based on the following: (1) including all firms, in particular those not covered by Thomson Reuters which therefore have zero news items by construction; (2) including only firms covered by Thompson Reuters. For each subset of models, a restricted and an unrestricted version of equation (2.7) is reported. The restricted version is the Poisson model, labeled "P", in which the variance of the gamma distribution is set to zero by imposing the restriction $\alpha = 0$. This version implies that the Poisson arrival rate is a non-stochastic function of trading activity based on a log-linear specification. The unrestricted version is the negative binomial model, labeled "NB", in which equation (2.7) is estimated without any restrictions on the parameters. For two firms with the same trading activity, this version allows the Poisson arrival rate to vary randomly according to the realization of a gamma distribution.

In table 2.7, results for the base case – where news items are counted only once and non Reuters firms are included in the data— appear in the first two columns. There are four important facts to note about the parameter estimates. First, the negative binomial model estimate of $\gamma = 0.68$ is almost identical to the estimate from the 30-bin model in table 2.5. This result is consistent with the intuition that the 30-bin model is an approximately unbiased way to model how news arrival rates vary with trading activity in a log-linear model. Second, the Poisson estimate of $\gamma = 0.81$ is much larger than the negative binomial estimate of $\gamma = 0.68$. This

result is consistent with the Poisson model producing biased estimates of the mean arrival rate of news events when expected arrival rates are not constant in the correct model. Third, the Newey-West standard error for γ is 0.024. This standard error is sufficiently large that the hypothesis $\gamma = 2/3$ is not rejected. It is, however, sufficiently small that the hypotheses of the alternative models, $\gamma = 0$ and $\gamma = 1$, are soundly rejected. Fourth, the Newey-West standard error of α or 0.218 is almost ten times smaller than the coefficient estimate $\alpha = 2.05$, indicating strong statistical support for the negative binomial model over the Poisson model.

In table 2.7, columns three and four give the results of Poisson and negative binomial models when the subset of firms is restricted to those in the Thomson Reuters dataset. The results have a flavor very similar to columns one and two. The coefficient estimate $\gamma = 0.65$ from the negative binomial model is again almost identical to the coefficient estimate in the 30-bin model from table 2.5. The Poisson estimate of $\gamma = 0.86$ is much higher than the estimate of the negative binomial model, indicating bias in the Poisson model. The standard error is still sufficiently small that the hypothesis $\gamma = \frac{2}{3}$ is not rejected, but sufficiently small that the alternative models are soundly rejected. The standard error of 0.120 on the coefficient estimate $\alpha = 1.63$ is so small that the model gives strong statistical support for the negative binomial model over the Poisson model.

In table 2.7, columns 5-8 provide estimates for both the Poisson and negative binomial models when news is counted by topic code, both for the dataset including "all firms" and the dataset including only firms in the "Thomson coverage" universe. When news is counted by topic code, the coefficient estimates of the negative

binomial, $\gamma = 0.71$ for "all firms" and $\gamma = 0.66$ for "Thomson coverage," are lower than the corresponding estimates of $\gamma = 0.74$ and $\gamma = 0.70$ for the 30-bin model in table 2.5. We conjecture that this difference arises because, conditional on a news item occurring, the count of topics for that news item follows a distribution that is different from a Poisson.

For example, consider firms for which the Poisson arrival rate is very low. Such firms are very likely to receive zero news items in a month, but occasionally receive one news item, and only rarely more than one. If it is common for news items to have more than one topic code, then the negative binomial model with news items counted by topic code will be trying to fit as a Poisson distribution a different distribution with too few cases of two or more events and not enough cases of one event.

This problem may be illustrated in figure 2.5. The top panel of figure 2.5 shows a plot of the frequency of $\log_2(N+1)$, where N is the number of news items and the result is rounded up to the nearest integer. Thus, letting x denote the value on the horizontal axis, the number $x = 0$ corresponds to $N = 0$, $x = 1$ corresponds to $N = 1$, $x = 2$ corresponds to $N = 2, 3$, $x = 3$ corresponds to $N = 4, 5, 6, 7$, $x = 5$ corresponds to $8 < N < 15$, $x = 6$ corresponds to $16 < N < 31$, etc. One observation consists of the number of news topics (for all news items) for one firm in one month. The bottom left panel shows the result obtained by subtracting the model frequencies for the Poisson model from the empirical frequencies. The positive value of about 15% for $x = 0$ shows that the data has a higher frequency of zero news articles than the Poisson model. The negative values for $x = 1, 2, 3, 4, 5, 6$

indicate that the data has a lower frequency of news articles for firms receiving fewer than 64 news articles per month. The bottom right panel shows the difference in frequencies obtained by subtracting the frequencies of the negative binomial from the frequencies in the data. The positive value for $x = 2$ indicates that the data has a higher frequency of $N = 2, 3$ than implied by the binomial model. The negative values for $x = 0$ and $x = 0$ indicate that the data has a lower frequency of $N = 0$ and $N = 1$ than the negative binomial model. Intuitively, this seems consistent with the hypothesis that when a news item occurs, there is a greater tendency for more than one topic code to be reflected than is consistent with negative binomial model structure, especially when the arrival rate of news items is low due to low trading volume. This issue is a subject for further research.

The last four columns of table 2.7 illustrate how estimates change when topic codes "RES" and "DIV" are eliminated from the data. Since RES represents news about earnings results and DIV represents dividend news, it might be conjectured that the presence of these news items are making the estimated coefficient for γ lower than when these items are omitted. This conjecture is consistent with the results, in that the estimated parameters with RES and DIV items omitted all greater than the corresponding results when RES and DiV are included in the counts. But note that excluding RES and DIV pushes the coefficient estimates for γ significantly above the hypothesis that $\gamma = 2/3$. Under the hypothesis that the true relationship is $\gamma = 2/3$, this means that Reuters journalists select the types of news items they cover based on the demand in the market for useful information. For firms with inactive trading, the selected news items tend to be dividend and earnings

announcements, but perhaps little else. For more actively traded firms, many other types of news items are reported, because the market demands it.

2.4.3 Robustness Checks

In table 2.7, we report standard errors based on the Newey-West procedure. As pointed out by Petersen (2009), it is possible that these standard errors are overstated. Month-level clustering and monthly fixed effects provide alternative ways to calculate standard errors. Table 2.8 reports results for four versions of the negative binomial model analogous to table 2.7, with panel A reporting coefficients and standard errors estimated with monthly clustering and panel B reporting coefficients and standard errors estimated with monthly fixed effects. For each of the fours versions of the model, the coefficient estimates for γ using monthly clustering are almost identical to the coefficient estimates using monthly fixed effects. Using both clustering and fixed effects, however, the coefficients estimates for γ are about 0.04 less than the estimates from table 2.7. For example, for the base model where each news item is counted once and all firms are included, clustering gives an estimate of $\gamma = 0.645$, fixed effects gives an almost identical estimate of $\gamma = 0.647$, and both estimates are less than the estimate of $\gamma = 0.68$ in table 2.7. As expected, the standard errors based on both clustering and fixed effects are lower than the standard errors Newey-West procedure. For example, in the base case, the standard error is 0.024 for Newey-West, while the standard error is 0.0143 for clustering and 0.0013 for fixed effects. Note that the standard error for clustering is about ten times larger than for fixed effects, indicating the some time series variation in coefficients may

be occurring.

The model of Kyle and Obizhaeva (2009) suggests that all variation in arrival rate of bets will be captured by variation in trading activity. This implies that if γ is held fixed at 2/3 in the following regression the coefficients $\beta 1$, $\beta 2$ and $\beta 3$ will be zero.

$$\mu = \left(\frac{W}{W^*}\right)^{\gamma} \times \left(\frac{V}{10^6}\right)^{\beta} 1 \times \left(\frac{P}{40}\right)^{\beta} 2 \times \left(\frac{\sigma}{0.02}\right)^{\beta} 3 \tag{2.9}$$

In upper panel of table 2.9, we provide Fama-MacBeth estimates for β_1, β_2 and β_3 from the monthly count data negative binomial model estimates where γ is held fixed at 2/3 and β_1, β_2 and β_3 capture sensitivity of arrival rate of news due to volume, price and volatility respectively. As in Kyle and Obizhaeva (2009), the scaling constant $W^* = (40)(10^6)(0.02)$ corresponds to the measure of trading activity for the benchmark stock with price \$40 per share, trading volume of one million shares per day, and daily volatility of 2%. Monthly estimates of β_1,β_2 and β_3 are averaged across the 72 months. The estimates are adjusted for auto-correlation up to 3 lags using the Newey-West procedure. F-statistics are calculated from the Fama-MacBeth regressions for three different models of trading activity. Estimates are obtained by six different sampling schemes. The lower 3 panels of table 2.9 report results of F-tests of the hypotheses $\gamma = \frac{2}{3}$ (i.e. $\beta_1 = \beta_2 = \beta_3 = 0$), $\gamma = 0$ (i.e. $\beta_1 = \beta_2 = \beta_3 = -\frac{2}{3}$) and $\gamma = 1$ (i.e. $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$) All estimates reject all three models of trading activity. All estimates are, however, closest to the hypothesis $\gamma = \frac{2}{3}$. Note that the negative binomial model implies that with $\alpha > 0$, other factors influence the flow of news items in addition to trading activity. The

results in table 2.9 are consistent with the interpretation that these other factors are correlated with volume, price and volatility.

In table 2.10 we estimate a quadratic effect of trading activity on news arrival rate by adding a quadratic term to the negative binomial model of table 2.7. The table presents estimates for sensitivity of the arrival rate of news articles to the square of trading activity controlling for trading activity itself. The errors are clustered at the monthly level. To avoid multi-collinearity, the level of trading activity is captured by $ln(W) - \mu$ and is constrained at 0.67, and the quadratic term is modeled as $(\log(W) - \mu)^2 - \sigma^2$. Here, μ is the sample mean of $\log(W)$ and σ is the sample standard deviation of $\log(W)$. This procedure ensures that the estimate for the constant is comparable with earlier models. Across all the models, the constant does not materially change by the inclusion of the quadratic term. The coefficient for the quadratic term, estimated to be between 0.026 and 0.043, is statistically significant since the standard error of 0.002 is less than one-tenth of the coefficient estimate.

In table 2.11, we report the quadratic effect and first order effect of trading activity on news arrival rate. Unlike table 2.10, we allow the effect of trading activity to be simultaneously estimated along with the quadratic effect. To avoid multi-collinearity the level of trading activity is captured by $ln(W) - \mu$ and the second order term is modeled as $(\log(W) - \mu)^2 - \sigma^2$. The estimates in table 2.11 for the quadratic term should not be different from the estimates reported in table 2.10 if the trading activity is not skewed. The estimates for the quadratic term are between 0.038 to 0.045. With the quadratic term, the estimates for γ ranges

from 0.55 to 0.61. If news items are counted only once, the estimate for γ is 0.57. Restricted to Thomson sample, in case where news items are counted only once, the γ in presence of a quadratic term is 0.55. If news items are counted by the topic code, the estimate for γ is 0.61. Restricted to the Thomson sample, if the news items are counted by the topic codes, the estimate of γ is 0.59. Note that in all cases, the estimate of γ in presence of the quadratic term is lower than reported earlier. It is, however, still closest to $\frac{2}{3}$, i.e. time clock irrelevance model.

It is possible that there are other factors like market capitalization, book to market and past performance that lead journalists to write more stories about a firm that does not reflect increase in information. One possible mechanism for higher market capitalization leading to larger number of articles is as follows. Large firms appear in the news as a placeholder for smaller firms. A journalist writing a story about a technology firm will like to place the firm in context of other technology heavyweights like Intel, Apple, Microsoft etc. These firms in turn will be mentioned, even though there is not much information about the large firm in the story. Similarly, firms with negative past performance are mentioned as journalist try to explain the decline in prices. These stories are likely to have no new information but for the past performance. For growth stocks the journalist might write stories since growth stocks have higher recognition among the reader base. Journalists are likely to be catering to a market of readers. The above-stated logic implies that large firms, growth firms and firms with negative price momentum may be more likely to have news stories in excess of the number predicted from trading activity (W).

Table 2.12 presents the Fama-MacBeth estimates for β_4, β_5 and β_6 from the monthly-count-data negative binomial model estimates where γ is held fixed at 2/3. β_4, β_5 and β_6 capture sensitivity of arrival rate of news due to previous month's market capitalization, book to market ratio, and past twelve months accumulated returns, respectively. The estimates are reported for the negative binomial model. We also examine whether the three coefficients are each different from zero, i.e. contain no information beyond what is already captured by trading time invariance theory. The coefficients for market capitalization and past performance are positive. To our surprise, however, an increase in book to market leads to more stories rather than fewer stories.

2.5 Conclusion and Implications

We take the Kyle and Obizhaeva (2009) model of trading games to data and find strong support for the model. The model is supported by data under various sampling scehemes. There are a couple of inferences to draw from this. First, the (Kyle and Obizhaeva 2009) model is based on the idea of cross sectional difference in trading days, a quantity that is not observable. On the other hand, empirical evidence presented in this paper strongly support the theory. Second, empirical research and theoretical modeling on how information is incorporated in to prices has focused on models where there is no cross sectional difference in trading time. Empirical support for (Kyle and Obizhaeva 2009) model suggests exploring cross sectional differences in trading time.

Popular approaches to study incorporation of information, focus on abnormal returns or change in volatility. The results in this paper suggest that cross sectional differences in trading time can provide rich ways to understand how information is incorporated into prices. The results also lend strong support to W as a reasonable measure of trading activity. It will be interesting to examine various factors that affect cross sectional variation in W.

2.6 Tables and Figures

Tab. 2.1: What makes news: List of topic codes

TOPIC	Brief Description	Number of news items	Proportion of news
STX	Regulations; Additions and deletions from indices; New listings; Delistings; Suspensions	508430	14.79%
RES	Corporate Results	475536	13.84%
MRG	Mergers and Acquisitions (including changes of ownership)	402443	11.71%
RESF	Corporate Results Forecasts	333921	9.72%
NEWS	Major breaking news	322301	9.38%
CORA	Corporate Analysis	228267	6.64%
DBT	Debt Markets	222902	6.49%
RCH	Broker Research and Recommendations	165252	4.81%
HOT	Hot Stocks	152253	4.43%
INV	Investing	131537	3.83%
REGS	Regulatory Issues	101693	2.96%
PRO	Biographies, Personalities, People	77476	2.25%
MNGIS	Management issues/policy	68819	2.00%
AAA	Ratings	51690	1.50%
IPO	Initial Public Offerings	30073	0.88%
PRESS	Press Digests	29795	0.87%
DIV	Dividends	28424	0.83%
JUDIC	Judicial processes/court cases/court decisions	26609	0.77%
WIN	Reuters Exclusive News	17829	0.52%
EXCA	Exchange activities	15061	0.44%
FED	Federal Reserve Board	12843	0.37%
ECI	Economic Indicators	11379	0.33%
BKRT	Bankruptcies	11166	0.32%
RSUM	Reuters summits	10243	0.30%
FES	Editorial special, analysis and future stories	267	0.01%
ERR	Error	204	0.01%
CFIN	Corporate Finance	143	0.00%
INSI	Technical Analysis	80	0.00%
CDM	Credit Market News	38	0.00%
TRN	Translated News	29	0.00%
CONV	Convertible Bonds	24	0.00%
NEWR	Original Corporate News Releases	1	0.00%

Table provides a listing of Topic codes for generating the news sample. Column 1 provides the topic code tag, column 2 provides a brief description. Column 3 provides the number of news items that were tagged with the particular topic code. A news item contains one or more topic code that indicate the broad subject of the item.

Tab. 2.2: Descriptive Statistics

Volume Groups:	All	1	2	3	4	5	6	7	8	9	10
	News counted only once										
Newscount	4.24	0.58	2.13	3.36	5.16	7.18	9.37	11.78	15.12	26.74	82.86
Newscount (median)	0	0	1	2	3	4	5	7	10	16	46
Newscount (maximum)	3344	143	183	242	221	198	367	259	817	1789	3344
Firms w/o any news in a month [a]	57.85%	72.95%	44.97%	35.16%	27.24%	21.71%	17.00%	13.33%	10.10%	6.59%	4.84%
Firms with one news item in a month	13.79%	15.11%	16.64%	13.62%	11.14%	9.26%	7.06%	5.99%	4.57%	2.43%	0.95%
Firms with >20 news item in a month	4.02%	0.07%	0.77%	1.77%	4.08%	7.91%	11.59%	17.19%	23.49%	41.27%	72.88%

Table reports the properties of news volume for all firms in our sample as well as for ten volume groups. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume.

[a] Expressed as a percentage of overall sample

Tab. 2.3: Descriptive Statistics

Volume Groups:	All	1	2	3	4	5	6	7	8	9	10
	News counted by topic code										
Newscount	9	1	4	6	9	13	17	21	27	47	145
Newscount_nodiv	8	1	3	5	8	11	14	18	23	41	131
Newscount_nodivnoradio	7	1	3	5	7	10	13	17	21	37	110
Newscount(Median)	1	0	2	3	5	6	9	13	17	28	84
Newscount_nodiv (Median)	0	0	1	2	4	5	7	10	14	23	74
Newscount_nodivnoradio (Median)	0	0	1	2	3	5	7	10	13	21	63
Newscount (Max)	7679	310	579	569	423	306	986	484	1407	3370	7679
Newscount_nodiv (Max)	7487	294	557	516	399	288	931	479	1316	3233	7487
Newscount_nodivnoradio (Max)	6727	244	496	441	374	252	840	402	1088	2523	6727
Firms w/o any news in a month [a]	39.0%	48.1%	34.6%	26.0%	19.2%	14.9%	11.3%	8.8%	6.6%	4.0%	2.2%
Firms with one news item in a month	7.6%	8.3%	9.1%	7.7%	6.3%	5.2%	4.4%	3.6%	2.8%	1.7%	0.6%
Firms with >20 news item in a month	6.5%	0.2%	2.4%	5.4%	10.9%	17.4%	24.0%	33.4%	42.0%	58.9%	81.7%
Average Volume($1000)	26361	1236	9051	17187	27428	39005	51348	68398	96258	156389	472866
Volatility	3.0%	3.3%	2.7%	2.5%	2.5%	2.4%	2.4%	2.3%	2.3%	2.3%	2.2%
W	653762	36380	241980	442557	691857	974793	1276220	1669274	2366748	3827421	11517954
Average Price	23.0	14.7	27.8	31.4	34.1	38.6	41.3	42.1	46.3	49.7	56.5
# of Observations	275059	166679	37170	15916	14715	7072	6730	6598	6583	6649	6947
	Includes firms without any news match										
Average Volume($1000)	21931	1028	8671	16634	26692	38218	50409	67315	95086	154837	465613
Volatility	3.1%	3.3%	2.7%	2.6%	2.5%	2.4%	2.4%	2.4%	2.3%	2.3%	2.3%
W	544846	30415	232598	428911	671684	954960	1252336	1646605	2343405	3784238	11355160
Average Price	21.1	13.6	27.1	30.8	33.7	38.1	40.9	41.7	45.9	49.2	55.8
# of Observations	340505	222543	41719	17620	16070	7622	7171	6947	6829	6841	7143

Table reports the properties of news volume and trading activity for two subsamples, the ones covered by Thomson Reuters and the one that are not covered. Panel A reports statistics for the ones covered by Thomson Reuters. Panel B reports statistics for all firms including the ones not covered by Thomson Reuters. Both panels report the daily dollar volume (in thousands of $), the average volatility of daily returns, the average price for all sample as well as for ten volume groups. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume.

Tab. 2.4: Proportion of firms with number of news items per month

Frequency	All news	All news but for dividend
0	48.24%	53.83%
1	9.46%	11.32%
2	9.53%	7.32%
3	4.26%	3.80%
4	4.60%	3.67%
5	2.54%	2.11%
6	2.37%	1.87%
7	1.55%	1.34%
8	1.58%	1.28%
9	1.14%	0.99%
10	1.08%	0.87%

Table presents frequency distribution for number of news per month as a proportion of total number of firms in the sample. Column 1 shows the number of news items per month. Column 2 shows the proportion of firms with n news items per month. Column 3 shows the proportion of firms with n news items per month when news items with topic code 'DIV' was not counted.

Tab. 2.5: Relationship between number of news items and trading activity

	News counted only once		News counted by topic codes	
	Thomson Coverage	All Firms	Thomson Coverage	All Firms
lnw	0.65 ***	0.68 ***	0.70 ***	0.74 ***
	0.005	0.005	0.005	0.005
R Square	0.89	0.90	0.90	0.91
F Value	227.55	267	251.79	309.12

Table presents estimates and standard error from log-linear models estimated across groups based on trading activity.

$$\log(newscount) = \beta_t \times d_t + \gamma \times \log(w) \tag{2.10}$$

The estimations is carried out with date fixed effect for log(number of news articles) as a function of log(W). All estimates suggest that γ is close to 0.67. ***: 1%, **: 5%, *: 10%,†: 15%

Tab. 2.6: Relationship between number of news items and trading activity across various topic codes

Topic Code	γ	se_γ	T test for γ=1 & p value		T test for γ=0 & p value		T test for γ=$\frac{2}{3}$ & p value	
AAA	0.57	0.01	-65.15	<0.001	86.47	<0.001	-15.12	<0.001
RES	0.45	0.00	-125.60	<0.001	103.42	<0.001	-50.03	<0.001
RESF	0.54	0.00	-98.80	<0.001	115.86	<0.001	-27.96	<0.001
DBT	0.66	0.01	-55.61	<0.001	108.90	<0.001	-1.32	19.1%
DIV	0.30	0.01	-106.65	<0.001	45.62	<0.001	-56.40	<0.001
HOT	0.72	0.01	-37.42	<0.001	96.99	<0.001	6.94	<0.001
RCH	0.67	0.00	-69.74	<0.001	142.36	<0.001	0.25	80.0%
MGIS	0.42	0.01	-86.48	<0.001	62.65	<0.001	-37.27	<0.001
MRG	0.60	0.01	-67.44	<0.001	99.93	<0.001	-12.21	<0.001
NEWS	0.95	0.01	-3.99	0.016%	77.07	<0.001	22.76	<0.001
PRESS	0.66	0.02	-19.44	<0.001	37.09	<0.001	-0.79	43.4%
PRO	0.68	0.01	-22.35	<0.001	47.34	<0.001	0.65	51.9%
REGS	0.52	0.01	-74.30	<0.001	79.53	<0.001	-23.53	<0.001
STX	0.75	0.01	-32.26	<0.001	98.84	<0.001	11.00	<0.001
# of Observations	72 cross sections of 30 trading group							
Method	Fama McBeth adjusted for autocorrelation							

Table presents estimates from log-linear models estimated for various topic codes. γ is the sensitivity of log(number of news items) to the log of trading activity. se_γ is the standard error of the estimate. The log linear model was estimated each month. Monthly γ was averaged across these month. The estimates are adjusted for auto-correlation up to 3 lags using the Newey-West procedure.
***: 1%, **: 5%, *: 10%,†: 15%

Tab. 2.7: Count data model: Relationship between number of news items and trading activity

	News counted only once				News counted by topic code				News counted by topic code no RES, DIV			
	All firms		Thomson Coverage		All firms		Thomson Coverage		All firms		Thomson Coverage	
	P	NB	P	NB	P	NB	P	NB	P	NB	P	NB
Constant	2.11	2.01	2.19	2.08	2.78	2.66	2.85	2.73	2.56	2.47	2.64	2.54
	0.044	0.036	0.037	0.028	0.049	0.037	0.044	0.030	0.052	0.043	0.045	0.035
γ	0.81	0.68	0.78	0.65	0.86	0.71	0.84	0.66	0.89	0.78	0.87	0.74
	0.007	0.024	0.007	0.018	0.008	0.025	0.010	0.019	0.007	0.025	0.009	0.020
Alpha		2.05		1.63		3.17		2.49		3.23		2.61
		0.218		0.120		0.325		0.170		0.318		0.185
-2 loglik	33179	14340	31443	13800	66498	17167	63140	16578	57697	15167	55084	14710
	748	384	862	415	1823	478	2038	510	1766	450	1925	477
BIC	33196	14366	31459	13825	66515	17192	63157	16602	57714	15193	55101	14735
	748	384	862	415	1823	478	2038	510	1766	450	1925	477

The table presents estimates for the arrival rate for the news article as a function of the trading activity. The arrival rate is modeled as $\mu = \left(\frac{W}{W^*}\right)^{\gamma}$. γ is deduced by modeling the news count as Poisson and Negative binomial process. The scaling constant $W^* = (40)(10^6)(0.02)$ corresponds to the measure of trading activity for the benchmark stock with price \$40 per share, trading volume of one million shares per day, and daily volatility of 0.02. The Poisson model (P) provides estimate for a Poisson arrival rate. Negative binomial model provides estimate for a Poisson arrival rate mixed with a Γ noise process. Monthly γ was averaged across these month. The estimates are adjusted for auto-correlation up to 3 lags using the Newey-West procedure. In each of the negative binomial model, the α is statistically different from 0, indicating Poisson model as an inferior modeling choice compared to a negative binomial model. Estimates for γ in the negative binomial model range from 0.65 to 0.78 depending on how the news is counted and what firms are considered as part of the sample.

Tab. 2.8: News arrival rate sensitivity to clustering and fixed effects

	PANEL A			
	News counted once		News counted by topic code	
	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage
Constant	1.995	2.071	2.639	2.718
	0.0215	0.0200	0.0229	0.0218
γ	0.645	0.611	0.661	0.624
	0.0143	0.0118	0.0153	0.0125
Cluster	Monthwise	Monthwise	Monthwise	Monthwise
	PANEL B			
Constant	2.086	2.049	2.846	2.779
	0.0262	0.0242	0.0296	0.0271
γ	0.647	0.616	0.670	0.633
	0.0013	0.0013	0.0013	0.001
Fixed Effect	Month	Month	Month	Month

Panel A of the table presents estimates for sensitivity of the arrival rate of news article to trading activity when the errors are clustered at monthly level. Panel B of the table presents estimates for sensitivity of the arrival rate of news article to trading activity when there are monthly fixed effects. The arrival rate is modeled as a negative binomial process. The models differ from one another in terms of sampling scheme.

Tab. 2.9: Robustness check for Volume, Volatility and Price

	News counted only once		News counted by topic codes		News counted by topic codes except DIV, RES	
	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage
Constant	2.19	2.14	2.91	2.84	3.10	3.03
	0.058	0.050	0.070	0.059	0.062	0.064
γ	0.67	0.67	0.67	0.67	0.67	0.67
$\beta 1$	0.08	0.06	0.09	0.07	0.20	0.18
	0.019	0.015	0.021	0.017	0.021	0.019
$\beta 2$	-0.22	-0.32	-0.17	-0.28	-0.19	-0.29
	0.032	0.017	0.034	0.018	0.032	0.018
$\beta 3$	-0.83	-0.84	-0.78	-0.81	-0.90	-0.92
	0.061	0.058	0.061	0.059	0.064	0.062
	Model of Trading Game Invariance: H_0*:* $\gamma = 2/3$					
$\beta 1 = \beta 2 = \beta 3 = 0$	177.6	500.0	126.4	367.3	190.0	342.6
	Model of Invariant Bet Frequency: H_0*:* $\gamma = 0$					
$\beta 1 = \beta 2 = \beta 3 = -2/3$	609.7	1082.4	526.9	922.4	841.8	847.1
	Model of Invariant Bet Size: H_0*:* $\gamma = 1$					
$\beta 1 = \beta 2 = \beta 3 = 1/3$	641.6	2065.5	464.8	1572.5	562.5	1122.3
-2 loglik	14041.8	13489.2	16923.7	16327.4	14886.1	14417.7

Table presents the Fama-MacBeth estimates $\beta 1$, $\beta 2$ and $\beta 3$ from the monthly count data negative binomial model estimates where

$$\mu = (\frac{W}{W^*})^{\gamma} \times (\frac{V}{10^6})^{\beta} 1 \times (\frac{P}{40})^{\beta} 2 \times (\frac{\sigma}{0.02})^{\beta} 3 \tag{2.11}$$

The scaling constant $W^* = (40)(10^6)(0.02)$ corresponds to the measure of trading activity for the benchmark stock with price \$40 per share, trading volume of one million shares per day, and daily volatility of 0.02. Negative binomial model provides estimate for a Poisson arrival rate mixed with a Γ noise process. The γ was fixed at 2/3. Monthly $\beta 1$,$\beta 2$ and $\beta 3$ were averaged across these month. The estimates are adjusted for auto-correlation up to 3 lags using the Newey-West procedure.F-statistics are calculated from the Fama-MacBeth regressions for three different models of trading activity.

Tab. 2.10: News arrival rate sensitivity to the square of trading activity I

	News counted once		News counted by topic code	
	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage
Constant	-0.17	0.27	0.46	0.92
	0.034	0.026	0.040	0.029
γ	0.67	0.67	0.67	0.67
Second Order	0.026	0.035	0.033	0.043
	0.002	0.002	0.002	0.0018
Cluster	Monthwise	Monthwise	Monthwise	Monthwise

The table presents estimates for sensitivity of the arrival rate of news article to the square of trading activity controlled for trading activity itself. The errors are clustered at monthly level. To avoid multi-collinearity the trading activity is constrained at 0.67 and the second order term is modeled as $(\log(W) - \mu)^2 - \sigma^2$. μ is the sample mean of $\log(W)$ and σ is the standard deviation of $\log(W)$. This procedure ensures that the three covariates - constant, trading activity, and quadratic term, are orthogonal to each other. The arrival rate is modeled as a negative binomial process. The models differ from one another in terms of sampling scheme.

Tab. 2.11: News arrival rate sensitivity to the square of trading activity II

	News counted once		News counted by topic code	
	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage
Constant	-0.03	0.40	0.51	0.96
	0.048	0.029	0.040	0.030
γ	0.57	0.55	0.61	0.59
	0.0132	0.010	0.014	0.010
Second Order	0.038	0.045	0.038	0.045
	0.0014	0.002	0.001	0.0014
Cluster	Monthwise	Monthwise	Monthwise	Monthwise

The table presents estimates for sensitivity of the arrival rate of news article to the square of trading activity when the trading activity is left as a free parameter. The errors are clustered at monthly level. To avoid multi-collinearity the trading activity is transformed to $\log(W) - \mu$ and the second order term is modeled as $(\log(W) - \mu)^2 - \sigma^2$. μ is the sample mean of $\log(W)$ and σ is the standard deviation of $\log(W)$. This procedure ensures that the three covariates - constant, trading activity, and quadratic term, are orthogonal to each other. This procedure ensures that the estimate for the constant is comparable with table 2.10. The arrival rate is model as a negative binomial process. The models differ from one another in terms of sampling scheme.

Tab. 2.12: News arrival rate in excess of trading activity

	News counted once		News counted by topic code	
	All firms	Thomson Reuters Coverage	All firms	Thomson Reuters Coverage
Constant	1.58	0.23	0.08	0.83
	0.159	0.332	0.476	0.344
γ	0.67	0.67	0.67	0.67
$\beta 4$	0.16	0.12	0.17	0.12
	0.028	0.022	0.031	0.023
$\beta 5$	0.26	0.2794	0.24	0.25
	0.025	0.026	0.021	0.023
$\beta 6$	-0.61	-0.62	-0.59	-0.61
	0.028	0.027	0.028	0.027
No information in excess of trading activity: $\beta 4 = \beta 5 = \beta 6 = 0$				
$\beta 4 = \beta 5 = \beta 6 = 0$	244.8	265.4	280.9	243.0
-2 loglik	10347	10010	12465	12096
	265.0	287.1	324.9	348.6
BIC	10387	10049	12506	12136
	265.0	287.2	324.9	348.7

Table presents estimates for sensitivity of the arrival rate of news article to market cap, book to market and past return in excess of the trading activity at the monthly level. The arrival rate is model as a negative binomial process. The models differ from one another in terms of sampling scheme. Monthly β was averaged across these month. The estimates are adjusted for auto-correlation up to 3 lags using the Newey-West procedure. The table also provides F statistic for the information from market cap, book to market and past return being jointly being equal to zero. The table shows that increased market cap, book to market and poor past performance lead to higher arrival rate for news articles. The F test rejects the coefficients jointly being equal to zero.

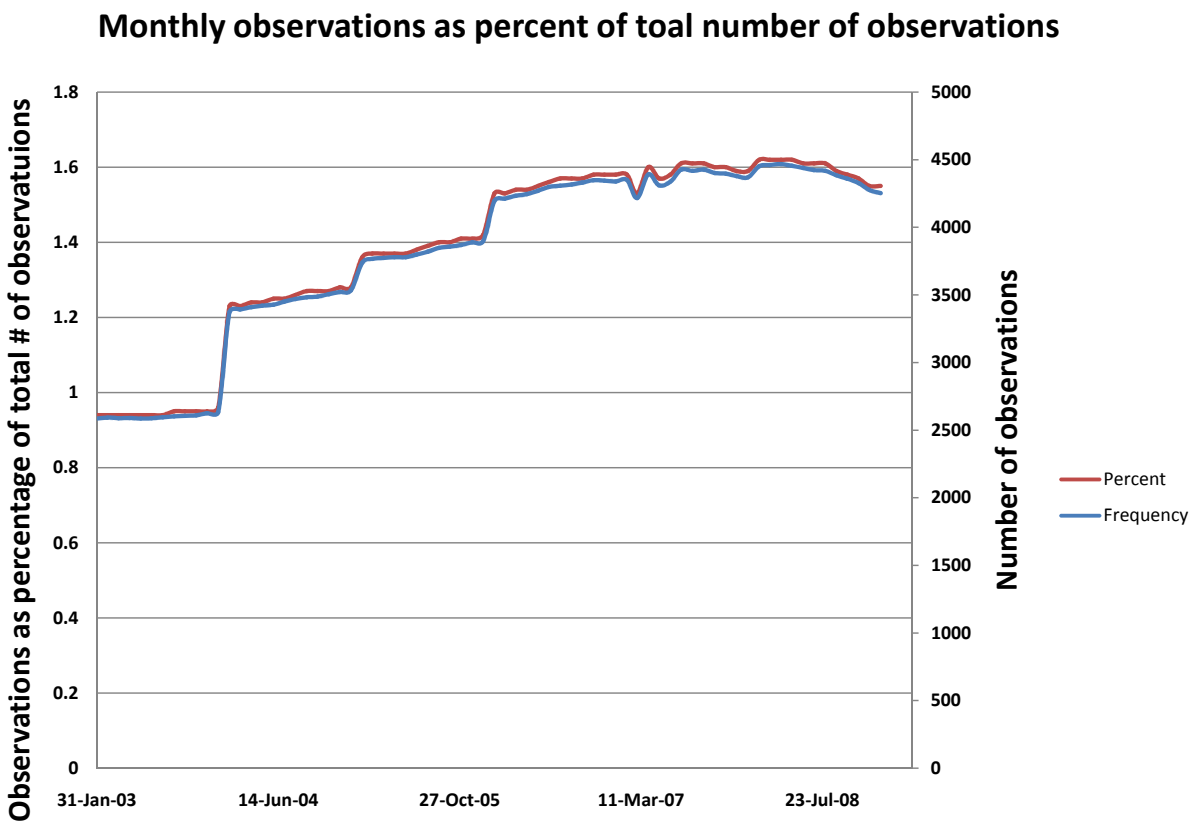

Fig. 2.1: Number of observations each month.

The number of observations per month is plotted as a percentage of total observations. The figure also shows the number of observations each month in the sample period. The chart shows that Thomson Reuters coverage of firms increased each year during our sample period.

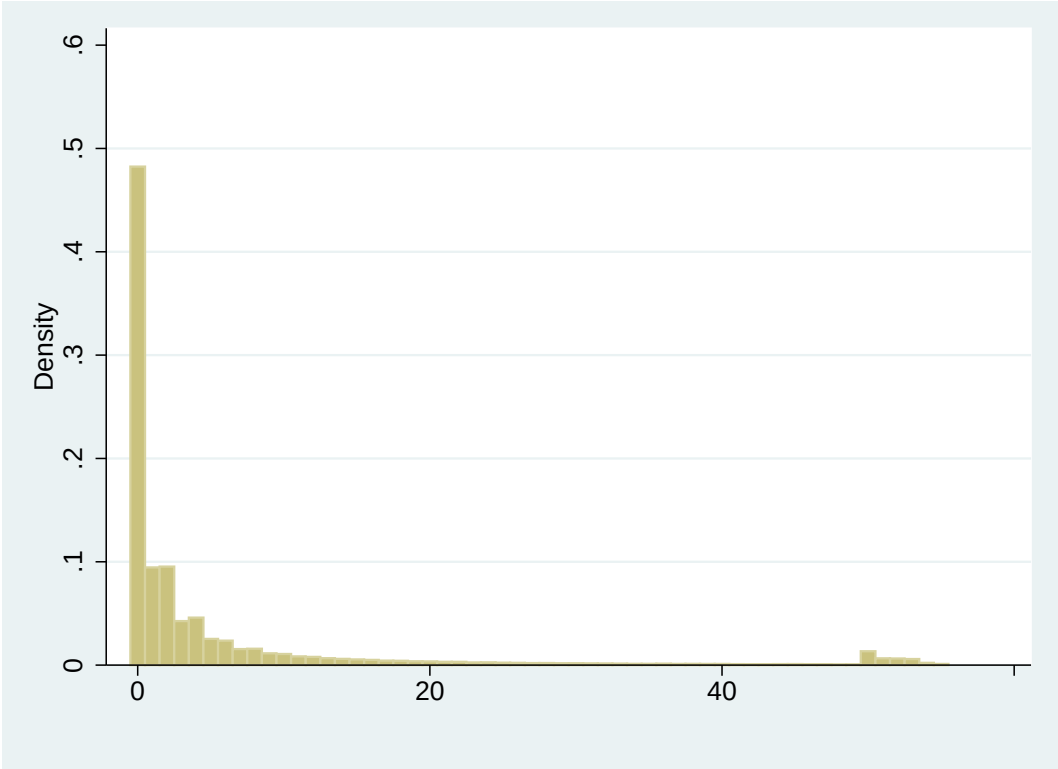

Fig. 2.2: Distribution of news items per month across firms

The figure shows the distribution of news items per month for our sample. Since there are some firms with large number of news items, the data is modified as follows. News items between 50 and 75 are recoded as 50, between 75 and 100 are recoded as 51, between 100 and 150 as 52, between 150 and 300 as 53, between 300 and 500 as 54 and more than 500 are coded as 55.

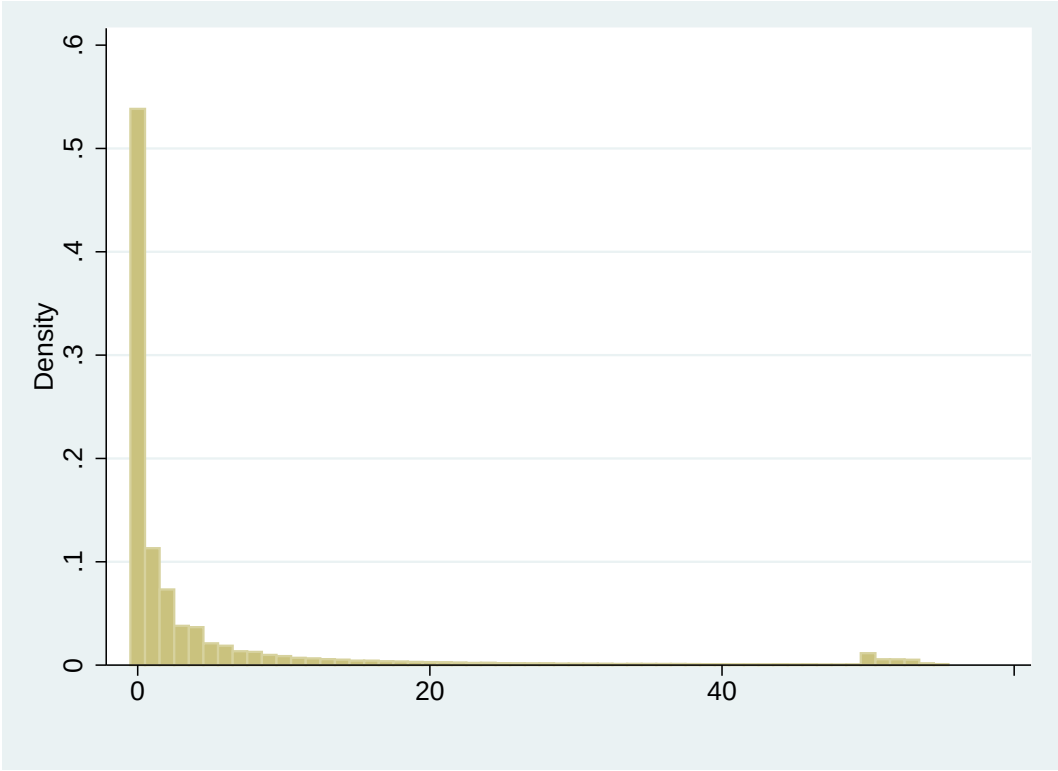

Fig. 2.3: Distribution of news items excluding dividends per month across firms

The figure shows the distribution of news items excluding dividends per month for our sample. Since there are some firms with large number of news items, the data is modified as follows. News items between 50 and 75 are recoded as 50, between 75 and 100 are recoded as 51, between 100 and 150 as 52, between 150 and 300 as 53, between 300 and 500 as 54 and more than 500 are coded as 55.

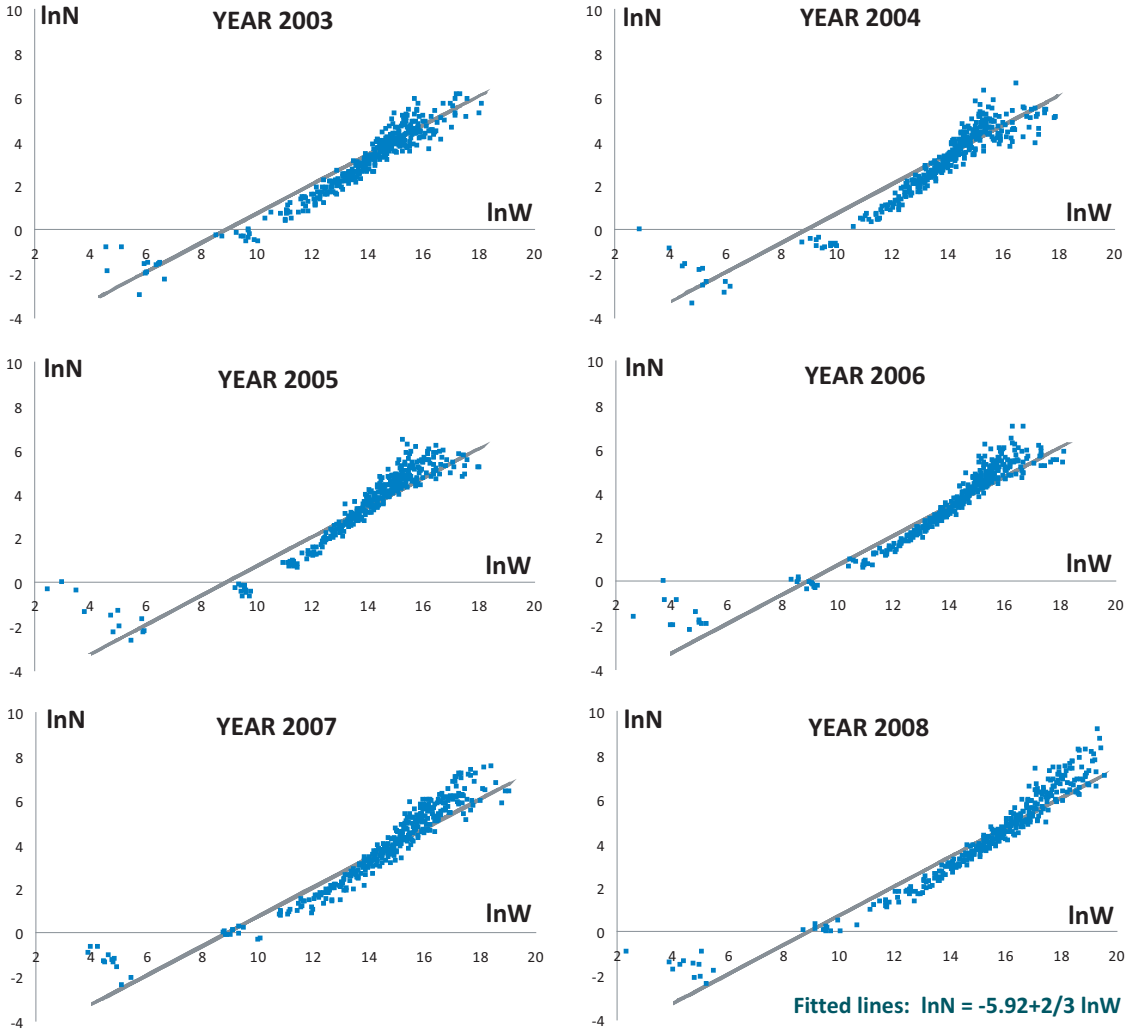

Fig. 2.4: log(Average number of news articles) plotted against log(average trading activity)

The figure shows how average news items varies with trading activity. First all stocks are grouped in thirty groups on the basis of increasing trading activity (W) such that each group has the same number of total news. The plot shows the average number of news articles every month as we move from one trading activity group to another for each of the six years of our dataset.

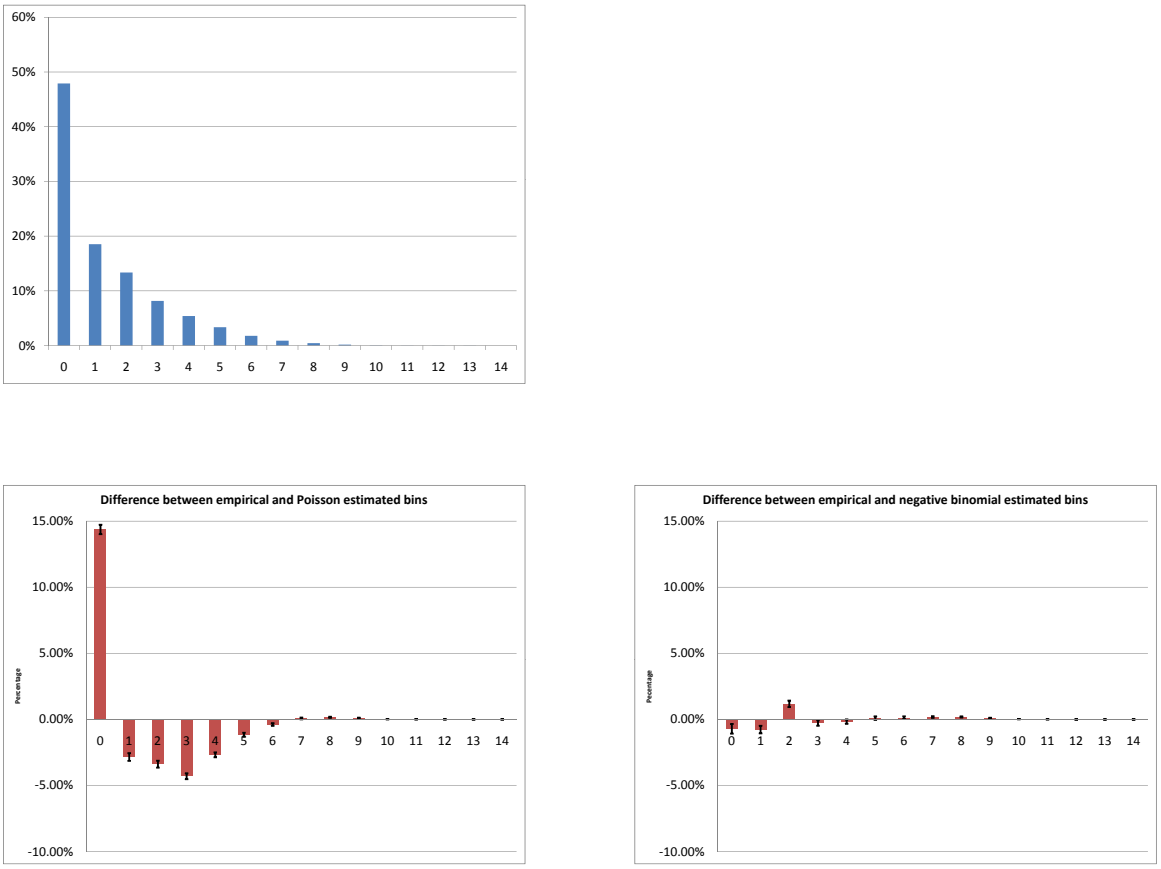

Fig. 2.5: Distribution of news per month across firms

The figure shows the distribution of news items for our sample as well as the difference between empirical distribution and estimated distribution for Poisson and Negative binomial models. The difference is calculated by subtracting estimated distribution from empirical distribution. The estimated distribution are estimated by using the Thomson Reuters coverage of firms. The news items are counted by the number of topic code mentions. Since there are some firms with large number of news items, the data is modified as follows. The bins are for $\log_2(Newsitems + 1)$. The bottom left panel shows the difference between empirical distribution and Poisson model. The bottom right panel shows the difference between empirical distribution and Negative binomial model. The figure shows that the Negative binomial model fits the data better than the Poisson model.

APPENDIX A: TEXT PROCESSING ENGINE

The sentiment engine has three major sequential processes: (1) pre-processing, (2) lexical and sentiment pattern identifier and (3) sentiment classifier.

Pre-processing: The sentiment engine pre-processes the text before attempting to ascribe any sentiment. The pre-processing consists of (1) sentence splitting, (2) tokenization, (3) part of speech tagging, (4) morphological stemming, and (5) shallow parsing. The sentiment engine first splits the whole news item into individual sentences. Sentences are further split into individual words, in computational linguistics literature this process is known as tokenization. Thereafter the sentiment engine identifies the parts of speech of each word. The classifier also morphologically stems the words, e.g. "gone", "went" and "goes" are all identified as "go". Morphological stemming identifies the root word for each word by matching each word to its root word. The sentiment engine does shallow parsing whereby it identifies the subject of the sentence and what is being said about the subject. For example, "The black cat slept" is identified as a Noun phrase relationship with the "The Black cat" as noun. The shallow parsing or identification of entity (or subject) for each sentence is used for providing relevance scores. The classifier keeps track of the subject for each sentence. The ability to keep track of the entity in discussion, allows the classifier to provide subject specific sentiment. If there are multiple subjects in a

sentence, it assigns different words to different subject. It also provides a relevance score for each subject.

Lexical and sentiment pattern identification: Followed by pre-processing, the sentiment engine does a lexical analysis identifying words into adjectives, adverbs, intensifiers, nouns and verbs. The lexical identification is important for sentiment processing since certain phrases and parts of speeches tend to convey sentiment information while others just aid in making a coherent sentence. The lexical identification feeds into sentiment patterns identification. Sentiment pattern consists of identification of negation, intensification, and verb resolution.

Sentiment Classifier: The sentiment classification is done by a three layer back-propagation neural network classifier with weight relaxation. Features extracted so far are used as input to the classifier. The classifier is trained using a random sample of triple annotated news articles spanning 14 months from December 2004 to January 2006. The annotation was done by analysts who analyze blogs and other outlets of public opinion. The annotation order was randomized, i.e., the manual annotators got the articles in random order and would not have been able to form long term opinions by reading the news articles. The whole process was supervised by a linguist and a former trader. The training articles are from the entire universe of Thomson Reuters coverage and include international stocks. The classifier produces three outputs between 0 and 1, which are probabilities of the article being positive, neutral or negative. The system achieves 75% accuracy against the average assesment of human analysts.

Many of the aforementioned methods are fairly standard in computer science

and are discussed at length in Manning and Schütze (1999). More details about the classifier are available in Infonic (2008).

BIBLIOGRAPHY

Akbas, Ferhat, Emre Kocatulum, and Sorin M. Sorescu, 2008, Mispricing following public news: Overreaction for losers, underreaction for winners, *SSRN Working Paper Series* pp. 1–44.

Amihud, Yakov, 2002, Illiquidity and stock returns: cross-section and time-series effects, *Journal of Financial Markets* 5, 31–56.

Ball, Ray, and Philip Brown, 1968, An empirical evaluation of accounting income numbers, *Journal of Accounting and Economics* 6, 159–178.

Barber, B.M., and T. Odean, 2008, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21.

Berry, T.D., and K.M. Howe, 1994, Public information arrival, *Journal of Finance* 49, 1331–1346.

Chan, W.S., 2003, Stock price reaction to news and no-news: Drift and reversal after headlines, *Journal of Financial Economics* 70, 223–260.

Chordia, T., S.W. Huh, and A. Subrahmanyam, 2007, The cross-section of expected trading activity, *Review of Financial Studies* 20, 709–740.

Clark, P.K., 1973, A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* 41, 135–155.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 1998, Investor psychology and security market under-and overreactions, *Journal of Finance* 53, 1839–1885.

Daniel, K., and S. Titman, 2007, Testing factor-model explanations of market anomalies, *Unpublished Working Paper* pp. 1–38.

Demers, E.A., and C. Vega, 2008, Soft information in earnings announcements: News or noise?, *FRB International Finance Discussion Paper No. 951.*

Dufour, A., and R.F. Engle, 2000, Time and the price impact of a trade, *Journal of Finance* 55, 2467–2498.

Engelberg, Joseph, 2009, Costly information processing: Evidence from earnings announcements, *Unpublished Working Paper, University of North Carolina Chapel Hill.*

Fama, E.F., and K.R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of financial economics* 33, 3–56.

Fang, Lily H., and Joel Peress, 2008, Media coverage and the cross-section of stock returns, *Journal of Finance* Forthcoming.

Foster, F.D., and S Viswanathan, 1994, Strategic trading with asymmetrically informed traders and long-lived information, *Journal of Financial and Quantitative Analysis* 29, 499–518.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H.E. Stanley, 2003, A theory of power-law distributions in financial market fluctuations, *Nature* 423, 267–270.

Graham, J.R., C.R. Harvey, and S. Rajgopal, 2005, The economic implications of corporate financial reporting, *Journal of Accounting and Economics* 40, 3–73.

Hong, H., and J.C. Stein, 1999, A unified theory of underreaction, momentum trading and overreaction in asset markets, *Journal of Finance* 54, 2143–2184.

Infonic, 2008, Reuters newsscope sentiment engine, Infonic Sentiment Technologies Whitepaper.

Jegadeesh, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Kayhan, Ayla, and Sheridan Titman, 2007, Firms' histories and their capital structures, *Journal of Financial Economics* 83, 1 – 32.

Kyle, A.S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Kyle, A. S., and A. A. Obizhaeva, 2009, Market microstructure invariants, University of Maryland.

Lesmond, D.A., M.J. Schill, and C. Zhou, 2004, The illusory nature of momentum profits, *Journal of Financial Economics* 71, 349–380.

Mandelbrot, B., and H.M. Taylor, 1967, On the distribution of stock price differences, *Operations Research* 15, 1057–1062.

Manning, Christopher D., and Heinrich Schütze, 1999, *Foundations of statistical natural language processing* (MIT Press).

Merton, R.C., 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483–510.

Mitchell, M.L., and J.H. Mulherin, 1994, The impact of public information on the stock market, *Journal of Finance* 49, 923–950.

Nieuwerburgh, Stijn Van, and Laura Veldkamp, 2009, Information acquisition and under-diversification, *Review of Economic Studies* Forthcoming.

O'Hara, M., 1995, *Market microstructure theory* (Wiley-Blackwell).

Peng, L., and W. Xiong, 2006, Investor attention, overconfidence and category learning, *Journal of Financial Economics* 80, 563–602.

Petersen, M.A., 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22, 435–480.

Sloan, R.G., 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *The Accounting Review* 71, 289–315.

Tetlock, P.C., M. Saar-Tsechansky, and S. Macskassy, 2008, More than words: quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

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

——— , 2009, All the news thats fit to reprint: Do investors react to stale information?, Unpublished working paper.

Valukas, Anton R. (ed.), 2010, *Chapter 11 Case Number 08-13555* vol. 1-9. United States Bankruptcy Court, Southern District of New York.