News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns

Steven L. Heston, and Nitish Ranjan Sinha*

April 23, 2014

Abstract

This paper uses a dataset of over 900,000 news stories to test whether news can predict stock returns. It finds that firms with no news have distinctly different average future returns than firms with news. We measure sentiment with the Harvard psychosocial dictionary used by Tetlock, Saar-Tsechansky, and Macskassy (2008), the financial dictionary of Loughran and McDonald (2011), and a proprietary Thomson-Reuters neural network. Simpler processing techniques predict short-term returns that are quickly reversed, while more sophisticated techniques predict larger and more persistent returns. Confirming previous research, daily news predicts stock returns for only 1-2 days. But weekly news predicts stock returns for a quarter year. Positive news stories increase stock returns quickly, but negative stories have a long-delayed reaction.

JEL–Classification: G12, G14

Keywords: News, Text Analysis

*Heston: Department of Finance, Robert H. Smith School of Business, University of Maryland, College Park, *sheston@rhsmith.umd.edu*. Ranjan Sinha: Board of Governers of the Federal Reserve System, *nitish.r.sinha@frb.gov*. We thank Tim Loughran, seminar participants at University of Maryland (finance brownbag), and Acadian Asset Management. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors.

1 Introduction

Textual information processing has become a growing part of empirical finance research. Tetlock's pioneering studies ((Tetlock, Saar-Tsechansky, and Macskassy 2008) and (Tetlock 2007)) demonstrated that news stories contain information relevant to predicting both earnings and stock returns. Subsequent studies have applied similar techniques with a variety of news sources, dictionaries, and methodologies. But the literature has not yet consolidated a uniform consensus of results.

Researchers have generally found that textual information can briefly predict returns. This occurs at the aggregate market level ((Tetlock 2007), (Dougal, Engelberg, García, and Parsons 2012), (Garcia 2013) and (Dzielinski and Hasseltoft 2013)) as well as the individual stock level ((Boudoukh, Feldman, Kogan, and Richardson 2013),(Sinha 2013) and (Chen, De, Hu, and Hwang 2014)). But the rapid growth of this empirical research has entailed the use of different dictionaries, datasets, and methodologies. For example, Tetlock, Saar-Tsechansky, and Macskassy (2008) used a general Harvard psychosocial dictionary, while Loughran and McDonald (2011) used a specialized financial dictionary. This may be appropriate in the sense that Tetlock used a broad sample of Wall Street Journal and Dow Jones News Service articles, whereas Loughran and McDonald used more specialized 10-K filings. Similarly, (Garcia 2013) used New York Times articles, whereas Jegadeesh and Wu (2013) also used 10-K's, Lerman and Livnat (2010) used 8-K's, and Chen et al (2014) used social media. These conflicting choices confound the type of source documents used for the textual analysis with the choice of dictionary. In particular, it begs the question of whether the improved dictionaries should be specialized to the source material, or whether more specialized textual processing can be effective for predicting stock returns based on a broad set of text source.

In addition to methodological differences, empirical studies have found substantively different types of predictability. Early work (Tetlock 2007) finds that short-term return predictability is quickly reversed at the market level. Loughran and McDonald (2011) find greater response for individual stocks within a multi-day event window. Garcia (2013) and Jegadeesh and Wu (2013) also find different results with market returns and individual stocks, respectively. In addition to aggregate market returns versus individual stocks, differences might stem from different source text, different dictionaries used to process the text, or other methodological choices. The duration and reversal of return predictability are important because they affect the economic interpretation of news in terms of permanent news impact or transient sentiment. As Tetlock (2007) summarizes, "The sentiment theory predicts short-horizon returns will be reversed in the long run, whereas the information theory predicts they will persist indefinitely".

This paper studies textual processing and compares return predictability of two different dictionaries with the return predictability of a sophisticated neural network. It applies these techniques on a large common set of Reuters news releases. This “apples to apples” comparison allows us to focus on the effect of different textual processing choices on return predictability. We confirm that the specialized Loughran and McDonald (2011) financial dictionary is superior to a general psychosocial dictionary for prediction of stock returns. Specifically, the specialized dictionary predicts larger returns that last longer and have less reversal than the general dictionary. And the neural network produces more predictability than either dictionary approach. This shows that different types of textual processing may extract different types of information from news stories.[1]

The duration of return predictability depends critically on the portfolio formation procedure. Previous research by Tetlock, Saar-Tsechansky, and Macskassy (2008), Loughran and McDonald (2011), and Lerman and Livnat (2010) has established a short-term response of stock prices to news. We find that stocks with positive (negative) news over one day have subsequent predictably high (low) returns for 1-2 days that are largely reversed. But in contrast to the published literature, aggregating news over one week produces a dramatic increase in predictability. Stocks with news over the past week have predictable returns for up to thirteen weeks. This is true even for stocks with only one news event per week. This shows it is important to gauge news sentiment relative to more than one day of stories.

Finally, our study controls for neutral news stories to isolate the effect of news tone on stock returns. We confirm the finding of (Fang and Peress 2009) that firms without news have different returns than firms with news. If no-news firms are compared to firms with news, then this effect can distort the comparison of positive news with negative news. Instead, we control for the effect of positive and negative news by comparing with neutral news. We find that news tone indeed has an effect on stock returns. Positive news predicts positive returns for only about one week, but negative news predicts negative returns for up to a quarter. This suggests that short sale constraints might limit the incorporation of information extracted by our textual processing techniques.

Section 2 describes the data and textual processing methods we use. Section 3 compares their ability to predict stock returns. Section 4 shows the existence of a

[1] Antweiler and Frank (2004) use Naive Bayes classifier to classify text. Das and Chen (2007) examines the effect of chat board messages on the stock prices using a voting scheme across multiple classifiers.

distinct news effect, and controls for this effect to contrast the different predictive pattern of positive and negative news sentiment. A final section concludes.

2 Comparison of Textual Processing Methods

The primary purpose of our study is to compare different methods of textual processing and interpret their ability to forecast individual stock returns. A distinguishing feature of our analysis is the use of multiple techniques of textual analysis on a common and broad dataset of news items. For example, Tetlock (2007) analyzes the Wall Street Journal's "Abreast of the Market" column, and Tetlock, Saar-Tsechansky, and Macskassy (2008) extended this to firm-specific stories in the Wall Street Journal and the Dow-Jones News Service. Loughran and McDonald (2011) used a more specialized list of financial words to analyze company 10-K reports. A natural question is whether the improvement in results from specialized processing will persist in a broad dataset, or whether it requires suitably specialized textual input. Therefore we use the "bag of words" dictionary approach and compare it with a neural network approach on identical data.

Another motivation for using a large, broad dataset is to increase power and distinguish different types of return predictability from different textual processing methods. For example, temporary market sentiment or news-induced trading liquidity should be quickly reversed. But new information should have a permanent impact on stock prices. Larger datasets and more powerful textual analysis methods have the potential to detect distinct patterns of predictability. In particular, we find that the temporal pattern of return predictability and reversal depends on the type of textual processing performed. We then interpret this in terms of permanent economic news versus transitory sentiment.

Finally, our dataset has a measure of the "tone" or sentiment of each news story.[2] This allows us to distinguish the effect of news publication from the effect of favorable or unfavorable news. The publication of news may draw attention to a stock, inducing both rational and irrational trading. This may affect the liquidity of the stock, and consequently change the expected return. We show that stocks with news have different expected returns from stocks without news. Controlling for this publication effect shows that positive and negative news are incorporated into stock prices at different speeds.

Our empirical analysis uses 900,754 articles tagged with firm identifiers from the Thomson-Reuters news system over the calendar years 2003-2010. Thomson Reuters provides a dataset of News sentiment called Thomson Reuters NewScope Data (sentiment data). This is a large dataset that is broader than many of the datasets previously studied. The dataset identifies the time of the news story (with millisecond resolution), the firm mentioned in the story, the headline of the news story, story id, the relevance of the news article for the firm, the staleness of a news item and measures from a neural-network-based sentiment engine. Thomson Reuters also provides a dataset called the Thomson Reuters news archive (text data) which contains the time of the news story, story id, the headline of the news story and the full text of the news item. We match the sentiment data with the text data using the timestamp and story id for all the items in the sentiment data and obtain a dataset that contains the text as well as the probabilities of the article being positive, negative and neutral.

Following Tetlock (2007), we use the Harvard General Inquirer Psychosocial Dictionary to measure the proportion of positive and negative words in each article, specifically the tag IV-4 positive and negative word lists. En-

[2]By contrast, the previous study of Akbas, Boehmer, Erturk, and Sorescu (2013) use the observed stock return to classify the tone of an article.

gelberg (2009) notes that the Harvard lists misclassify positively many terms that are neutral financial meaning, such as company, shares, and outstanding. Loughran and McDonald (2011) similarly note that the Harvard negative list erroneously contains many terms that are neutral in a financial context, such as (tax, costs, capital, cost, and expenses). Therefore we also use Loughran and McDonald's positive and negative financial word lists available at `http://www.nd.edu/~mcdonald/Word_Lists.html`.

In addition to the Harvard and Loughran-McDonald word lists, we use a neural-network-based sentiment engine to classify articles as positive, negative, or neutral.[3] The Thomson Reuters sentiment engine consists of three stages - pre-processing, lexical and sentiment pattern identification, and the neural network. The basic unit of analysis is a sentence. First the document is segmented into individual sentences. Then, at the pre-processing stage, the subject (company) of the sentence is identified followed by what is being said about the subject (company). Some stories mention multiple companies. Identifying the subject for each sentence allows for different sentiment for different firms mentioned in the same story. At the pre-processing stage the classifier maps each word to corresponding parts-of-speech. Typically in English language parts-of-speech such as preposition and articles are not used to denote the sentiment. At the lexical and sentiment pattern identification stage, some of the parts-of-speech are retained and fed into the proprietary neural network classifier. The neural network is trained by 3000 randomly selected news stories which were tagged by three former traders. The neural network accounts for word order, modifying adjectives and common phrases in the finance parlance.

To illustrate the differences across these methods, consider the following fa-

[3]The Ravenpack database used by von Beschwitz, Keim, and Massa (2013) also uses sentiment analytics.

vorable article.

> Fitch may raise J.C. Penney ratings, NEW YORK, June 17 (Reuters) - Fitch Ratings Services may raise J.C. Penney Co. Inc. <JCP.N> senior unsecured notes citing improved operating trends at its department stores and potential debt reduction from the sale of of its Eckerd drugstore business. Fitch said it may raise J. C. Penney's "BB" senior unsecured notes, it "BB-plus" secured bank facility and its "B-plus" convertible subordinated notes. The action affects $5.2 billion of debt. "Penney continues to make solid progress in turning around its department stores and catalog/internet business. The segment's comparable store sales increased a strong 9.4 percent in the four months ended May 2004, and have been positive for three consecutive years," said Fitch.

According to the neural network classifier, this article is strongly positive with an 85% chance of being positive, only 3% chance of being negative, and remaining 12% chance of being neutral. The Harvard dictionary finds three positive words: "progress", "positive", and "make." But it finds four instances of negative words, consisting of three instances of "raise" and also the word "make." Curiously, the word "make" appears on both the Harvard positive and negative lists. Hence, the Harvard dictionary assigns this article a negative net rating. The financial dictionary produces different results. It finds four positive words, "strong", "improve", "progress", and "positive", with no negative words. In this case the financial dictionary provides a sensibly positive assessment by correctly identifying these words used in a positive context.

But the opposite case also occurs. The neural network finds the following story about Bel Fuse to be as positive as the previous story about J.C. Penny. It also rates 85% positive, 3% negative, and 12% neutral.

> Bel Fuse says Technitrol's raised offer undervalues stock, April 24 (Reuters) - Network equipment maker Bel Fuse Inc. <BELFA.O> said Technitrol Inc's <TNL.N> increased offer to buy it for $43 per share, like an earlier offer, significantly undervalues its stock. In a statement, Bel confirmed that it is engaged in the bidding stage with respect to one potential acquisition. However, it declined to provide further details.

The Harvard dictionary also rates it favorably, having eight positive words and no negative words in a very short article. But in this case, the Harvard dictionary appears to get the right answer for the wrong reasons. The positive words were three occurrences of "offer", along with one occurrence of "share", "buy","significant", "engage", and "respect." In contrast, the financial dictionary list finds no positive words, and only one negative word, "declined". In this case the Harvard psychosocial dictionary is closer to the neural network.

While these articles illustrate potential pitfalls of a "bag of words" approach, it is not feasible to individually examine all articles in the database. So Table 1 presents summary statistics for our text measures. Panel A show that the 5.3% average proportion of positive words in the average article outnumbers the 2.9% proportion of negative words in the Harvard list. But the financial list has a reverse bias, 2.9% negative and only 0.5% positive. The neural network has a different scale, predicting an average 29.9% chance that an article is positive, and 27.5% chance that an article is negative.

===Insert Table 1 here====

We begin the analysis by measuring the "tone" of an article as the net proportion of positive words in excess of negative words. In the case of the neural network, we subtract the probability of a negative rating from the probability of a positive rating. This gives us net measures of the favorability of an article about a company (Section 4 examines the positive and negative word lists separately). Table 1, Panel B, shows that the Harvard dictionary rating is actually negatively correlated with the financial dictionary rating, with a nontrivial correlation of -22%. It appears that the presence of net positive general words in related to the net absence of positive financial words. The Reuters measure of net sentiment is reassuringly positively correlated with both dictionary measures. Table 1, Panel C, breaks the correlations down into positive and negative components. It shows another curious feature of the general Harvard list - the proportion of positive words has 51% correlation with the proportion of negative words in the Harvard list, and 53% correlation with negative words on the financial list! For the financial dictionary and the neural network, the respective correlations between positive and negative words are a more sensible -23% and -42%. This may help explain why Tetlock (2009) and Loughran and McDonald (2011) found most predictability with negative words. Panel C breaks down the correlations separately among the positive and negative ratings. The proportion of positive words in the Harvard dictionary is negatively correlated with the proportion of positive words in the financial dictionary, and also with the positive rating of the neural network. The Reuters neural network negative rating is only very weakly correlated with the proportion of negative words (9% with Harvard dictionary, 1% with financial dictionary). Clearly these textual approaches produce dissimilar conclusions.

===Insert Table 2 here====

Table 2 shows summary statistics for firms sorted by size. Firms in the largest decile have frequent stories, 7.63 stories per week exceeds on per day. But the small firms are comparatively neglected in news coverage. Firms in the smallest two deciles average less than one article per week, and over 90% of firms in the smallest three size deciles receive no news in a given week. Over the ten-year sample period, small firms underperformed large firms; the smallest decile lost 0.14% per week, while the largest decile gained 0.06% per week. The last column reveals a critical feature relevant to our study of textual analysis. Firms that receive news coverage in a given week have different average returns than typical firms of their size in the subsequent week. Note that these "no news" returns are not subject to a survivorship bias or short-term informational effect, as in "the dog that didn't bark". The most dramatic difference occurs in the smallest decile, where the average small firm lost 0.14% in a given week, but small firms with news averaged 2.00% in the week following the news. Given that small firms are often illiquid and costly to trade, this may not represent a profit opportunity. But it documents that firms with news are distinctly different from firms without news. When measuring the effect of news sentiment, it will be important to control for the existence of news.

3 Predicting Returns

Our simplest test of news sentiments uses portfolios based on net sentiment, positive minus negative. In contrast to previous studies which used SEC filings or periodic newspaper columns, our dataset has almost one million news stories, sometimes with multiple stories about a particular firm. Therefore we measure the sentiment for a given firm as the average positive minus negative sentiment on all stories about that firm in a formation period. Table 3 presents excess returns on quintile spreads, i.e., the difference between returns on the highest and lowest sentiment portfolios using the Harvard psychosocial dictionary, the Loughran-McDonald financial sentiment dictionary, and the Thomson-Reuters sentiment engine. The quintiles are formed daily on Day 0, and returns are reported daily. We use quintiles instead of deciles or more selective portfolios because many stocks do not have news results on any given day.

The contemporaneous returns on the Day 0 news release show that more sophisticated textual processing producer larger returns, both economically and statistically. The Harvard dictionary has an associated publication day (excess) return of 0.61%, the Loughran-McDonald dictionary has a higher return 1.21%, and the Thomson-Reuters method has the highest return of 1.99%.

===Insert Table 3 here====

Note that average excess returns on the quintile spreads are invariably positive in the ten days preceding the publication of news, usually with t-statistics exceeding 2. This is expected, since news stories may lag events that affect stock prices.

The more interesting result is the post-publication returns. The Harvard dictionary does not successfully predict positive post-publication returns. But the Loughran-McDonald dictionary produces an average return of 0.08% with a t-statistic of 4.1 on Day 1, and the Thomson-Reuters method produces returns of 0.17% on Day 1 and 0.04% on Day 2, both significant at the 95% level. It appears that the more complex methods of textual processing also predict stock returns that are larger and more permanent than returns predicted by simpler methods.

===Insert Figure 1 here====

Figure 1 contrasts these methods by showing cumulative daily excess returns for one quarter, or 63 business days after the news date. The positive announcement returns to the Harvard quintile spread are largely reversed by the end of quarter. The Day 1 predictability of the Loughran-McDonald quintiles is reversed by the end of quarter, but the Day 0 announcement return remains. For the Thomson-Reuters method, the Day 0 and Day 1 returns are not reversed, and the subsequent performance is rather flat. The predictability of different processing methods differs both quantitatively and qualitatively.

These daily results are consistent with previous findings with daily data. Tetlock, Saar-Tsechansky, and Macskassy (2008), Loughran and McDonald (2011), and Lerman and Livnat (2010) all find predictability over event windows of 1-4 days, with varying degress of reversal. This previous research used periodic news columns and SEC filings. Those datasets typically have only one news item per firm. In contrast, our dataset often has multiple news stories about firms spread over several adjacent days. Given our dataset with frequent stories, daily aggregation might not be the best choice.

===Insert Table 4 here====

Table 4, panel A shows that the predictability changes dramatically when quintile portfolios are formed based on weekly news.[4] The Harvard dictionary produces a quintile spread return of 1.21% in Week 0; the Loughran-McDonald dictionary yields 2.20%, and the Thomson-Reuters method produces 3.75%.

[4] We found similar results with biweekly and monthly aggregation.

These numbers must be interpreted with caution, since the news stories may be published subsequent to a day on which news actually caused high returns. But the subsequent weekly returns are truly out-of-sample. The Harvard dictionary does not predict any significant positive weekly returns, and several of the point estimates are slightly negative. The Loughran-McDonald dictionary produces a quintile spread return of 0.33% in the week following news publication, with a t-statistic of 4.3. While the subsequent returns are not statistically significant in any individual week over the next quarter, all but one are positive. The best predictability comes from the Thomson-Reuters method. It predicts positive returns for 13 weeks after the new story release. Most of them are statistically significant at the 95% level, including a 0.21% return in Week 13. Figure 2 graphs the cumulative returns from these three weekly strategies. It shows the clear ordering of predictability across the three methods.

===Insert Figure 2 here====

There are two explanations for the striking improvement in predictability when using weekly returns. One explanation is that some firms have multiple news stories over different days within a week, and the predictability stems from the information confirmation of these clustered news stories. Of the firms with news in a given week, only 35% have more than one news story, and only 9% have more than two. These firms with multiple news stories tend to be larger than firms with less news coverage. This explanation says that this minority of firms drives the profitability of weekly strategies.

===Insert Figure 3 here====

A second more prosaic explanation is that daily news is quite volatile. Figure

3 illustrates this by graphing the 20th and 80th percentiles of Thomson-Reuters news sentiment based on daily and weekly news. It is clear that the thresholds for daily quintile sentiment are quite volatile.[5]. Some days simply have little news, or little news with strong sentiment, while others have an abundance of news with strong positive or negative sentiment. At a few points, the daily 20th and 80th percentile lines almost touch. This means that firms with stories in the highest quintile of sentiment on one day would be in the lowest quintile on an adjacent day! Clearly this is a noisy way of classifying firms based on sentiment. The weekly cutoffs still show some variation, but are much more stable through time.

===Insert Figure 4 here====

Table 4, Panels B and C, separately reports the weekly quintile spreads for subsamples of firms that have one and multiple news stories in Week 0. Both subsamples are profitable in twelve of the thirteen post-news weeks. The quintile spread of firms with multiple news stories is positive at the 95% level of significance in Week 11, whereas the quintile spread of firms with a single weekly news story is significantly positive at the 95% level in Week 13. Figure 4 graphs the cumulative returns to these strategies over thirteen weeks. It shows the quintile strategy based on firms with multiple weekly news items is more profitable over the quarter, but both subsamples generate a strong, profitable trend. This is quite different from the flat daily results in Figure 1.

[5]One can also see the news thresholds shrank after the Thomson-Reuters database was expanded in 2005

4 News and Positive Versus Negative Sentiment

The previous results show a delayed reaction to weekly news about firms. Specifically, firms with good news over a one-week subsequently outperform firms with bad news over a one-week period. Portfolios formed on this basis earn excess returns for up to 13 weeks.

But this exercise does not completely disentangle the "news effect" from the sentiment effect. It is conceivable that the mere publication of news about a firm affects its returns, regardless of the content or sentiment of the news. For example, a news article with little new information might nevertheless make its information common knowledge. This could resolve information asymmetry and thereby change the liquidity of a market. Like the "dog that didn't bark", the mere fact that articles were published or not published about a firm contains information.

Another limitation of the quintile spread results is that they do not reveal whether the predictability stems from positive or negative news. For example, Tetlock (2009) and Loughran and McDonald (2011) found that the preponderance of return response stems from negative news. In addition, the magnitude of a "news effect" is necessary to guage the distinct effects of positive and negative news. In order to compare the potentially dissimilar effects of positive and negative news, we need to compare to an appropriate return benchmark. The summary statistics in Table 2 show that firms without news have underperformed firms with news over our sample period. Inclusion of those firms using a portfolio methodology would bias the relative comparison of firms with different types of news. Specifically, firms with news would appear to outperform firms without news, regardless of the sentiment of the news, and that would exaggerate the impact of positive news while reducing the apparent effect of negative news. In order to distinguish a publication effect from the quality of the information, and in order to separately evaluate the effect of positive and negative news, we must use a multivariate technique. This technique must separate measure the news effect, the effect of positive sentiment in news, and the effect of negative sentiment in news.

We use the cross-sectional regression technique of Fama and MacBeth (1973). For a given lag k ranging from 0 to 13, we regress stock returns on sentiment ratings

$$r_i(t) = \alpha_k(t) + \gamma_k(t) * 1_{If_{news}}(t-k) + \beta_k(t) * Positive_i(t-k) + \delta_k(t) * Negative_i(t-k) + \epsilon_i(t) \quad (1)$$

where $r_i(t)$ is the return on stock i in week t, $1_{If_{news}}$ is a dummy variable for firms with news over the given lag, and $Positive_i(t-k)$, $Neutral_i(t-k)$ and $Negative_i(t-k)$ are the evaluation of sentiment in news articles published in the lagged week. Following Fama (1976), we can interpret $\alpha_k(t)$ as the return on an equally weighted portfolio of firms with no news at lag k. If "no news is good news," then α_k will tend to be negative. However, the summary statistics show that firms without news tend to underperform, so we expect this intercept to be positive. The term γ_k represents the return premium for firms that have neutral published news over firms with no news. The β_k's and δ_k's represent excess returns on costless, well-diversified portfolios that have 100% net loadings on positive or negative sentiment variables at a given lag.

===Insert Table 5 here====

Table 5 presents the average results of the regression coefficient time-series, along with time-series t-statistics using the Harvard dictionary, the Notre Dame dictionary, and the Reuters sentiment engine. The "no news" intercept is identical in all regressions, so we only report it in the first Panel. It is negative at all lags, ranging from -1 basis point per week to -6 basis points per week. While these average returns are not statistically significantly different from zero, this shows that firms without news have performed poorly over the sample period.

The premium for neutral news is $\gamma_k(t)$. It represents the weekly return premium of firms with 100% neutral news over firms with no news. The average point estimates are positive at all nonzero lags, showing there is a positive effect of neutral news. This contradicts the well-known adage that "No News is Good News" popularized by Campbell and Hentschel (1992). Interestingly, the contemporaneous lag 0 estimate for the Loughran McDonald financial dictionary has a highly significant t-statistic of 7.0. Recall from the summary statistics in Table 2 that only a small percentage of words appear in this financial dictionary. It appears that news stories without positive or negative financial terms are associated with an immediate positive return response. While the lagged effect of neutral news is insignificant for the Harvard dictionary, it is statistically significant at the 95% level for a number of weekly lags in the case of the Notre Dame dictionary and the Reuters engine. Rather than "No news is good news," the lesson is that "Neutral news is good news."

If neutral news is good, then we should expect that positive news is even better. Table 5 confirms this. The contemporaneous (lag 0) β_0 effect is positive and highly statistically significant for all measures of sentiment. If news travels slowly, then good news should also have a positive lagged effect. But this does not appear to be the case. The Harvard point estimates are not statistically significant at any lags. The Loughran McDonald financial dictionary and Thomson Reuters estimates are marginally significant at the 95% level at the first weekly lags, but are not statistically positive at further lags. The subsequent point estimates are near zero and have different signs at higher lags. It appears that the market quickly incorporates positive information into returns.

Negative news also has a strong immediate effect on returns for all three measures of sentiment. It does not seem to have a lagged effect for the Harvard measure. The Loughran McDonald measure of negative sentiment also affects returns at the first weekly lag, but is not significant thereafter. But there is a strong lagged effect for Reuters measure. This influence of Reuters sentiment is negative at all 13 lags, and the individual weeks are statistically significant at the 95% level at lags 1-6 and lag 10. This echoes Hong, Lim, and Stein (2000) finding that "Bad news travels slowly." The findings are, however not consistent with Pound and Zeckhauser (1990) who find that rumor is already incorporated into prices.

===Insert Figure 5 here====

Figure 5 graphs cumulative coefficients from Table 5 for horizons ranging from one week to thirteen weeks. This illustrates the pattern of news impact over different time periods. It shows that the effect of neutral news is small, but accumulates positively for a full quarter. The incremental effect of positive (Loughran-McDonald) sentiment is only positive for two or three weeks, and then flattens out to negligible levels. In contrast, the impact of negative sentiment continues to be strong for the full thirteen week period. Overall, neutral news, positive news, and negative news have different patterns of predicting stock returns through time. These findings demonstrate that is important to carefully measure and distinguish news sentiment. It reinforces the portfolio results which show that more detailed financial measures and neural network parsing help predict stock returns. The persistent predictive ability of negative Reuters sentiment is interesting in this regard. It is consistent with short sale constraints that prevent a small informed minority from fully impacting stock prices. The lack of lagged predictability of dictionary approaches suggests that easily digested information is more rapidly incorporated into prices.

5 Conclusion

This paper compares different types of textual processing in their ability to predict stock returns. We compare the dictionary approach of Tetlock (2009) and Loughran and McDonald (2011) with a neural network approach applied to a broad dataset of news stories. This establishes successful application of dictionary approaches to a new dataset and shows that dissimilar textual approaches can succeed on a common dataset. A simple correlation analysis shows that different dictionary and neural network measures provide very different evaluations of news sentiment. Importantly, the sign, magnitude, and duration of stock return predictability depend on the type of textual processing and horizon of news aggregation.

Using the more relevant Loughran-McDonald financial dictionary or a neural network approach increases the magnitude and horizon of return predictability. Most predictability is quickly reversed when news is measured over a daily period. But predictability lasts for up to a quarter when news is aggregated over a week. This shows the importance of properly aggregating news to measure its effect on stock returns. It also indicates that the effect of news on prices is not only due to transient sentiment or liquidity. Deeper textual measures based on weekly news appear to detect news that is persistently under-incorporated into current stock prices.

This paper also distinguishes the effect of news from the positive or negative sentiment of that news. There is a news effect, where firms with neutral news outperform firms without any news. Controlling for the news effect shows that positive news affects stock prices within one week. But negative news predicts low stock returns for up to one quarter.

References

Akbas, F., E. Boehmer, B. Erturk, and S. M. Sorescu, 2013, "Short Interest, Returns, and Fundamentals," *SSRN Working Paper Series*, pp. 1–51.

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

Boudoukh, J., R. Feldman, S. Kogan, and M. Richardson, 2013, "Which news moves stock prices? A textual analysis," *NBER Working Paper No. 18725.*

Campbell, J. Y., and L. Hentschel, 1992, "No news is good news: An asymmetric model of changing volatility in stock returns," *Journal of financial Economics*, 31(3), 281–318.

Chen, H., P. De, J. Hu, and B.-H. Hwang, 2014, "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media," *Review of Financial Studies*, 27(5), 1367–1403.

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

Dougal, C., J. Engelberg, D. García, and C. A. Parsons, 2012, "Journalists and the stock market," *Review of Financial Studies*, 25(3), 639–679.

Dzielinski, M., and H. Hasseltoft, 2013, "Aggregate News Tone, Stock Returns, and Volatility," Unpublished Working Paper.

Engelberg, J., 2009, "Costly Information Processing: Evidence from Earnings Announcements," *Unpublished Working Paper, University of North Carolina Chapel Hill.*

Fama, E. F., 1976, *Foundations of Finance: Portfolio Decisions and Securities Prices*. Basic Books.

Fama, E. F., and J. D. MacBeth, 1973, "Risk, return, and equilibrium: Empirical tests," *Journal of Political Economy*, 81(3), 607–636.

Fang, L. H., and J. Peress, 2009, "Media coverage and the cross-section of stock returns," *Journal of Finance*, 64(5), 2023–2052.

Garcia, D., 2013, "Sentiment during recessions," *Journal of Finance*, 68(3), 1267–1300.

Hong, H., T. Lim, and J. Stein, 2000, "Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies," *Journal of Finance*, 55(1), 265–295.

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

Lerman, A., and J. Livnat, 2010, "The new form 8-K disclosures," *Review of Accounting Studies*, 15(4), 752–778.

Loughran, T., and B. McDonald, 2011, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 66(1), 35–65.

Pound, J., and R. Zeckhauser, 1990, "Clearly heard on the street: The effect of takeover rumors on stock prices," *Journal of Business*, 63(3), 291–308.

Sinha, N. R., 2013, "Underreaction to News in the US Stock Market," University of Illinois at Chicago Working Paper.

Tetlock, P., M. Saar-Tsechansky, and S. Macskassy, 2008, "More than words: quantifying language to measure firms' fundamentals," *Journal of Finance*, 63(3), 1437–1467.

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

——— , 2009, "All the News ThatŚs Fit to Reprint: Do Investors React to Stale Information?," Unpublished working paper.

von Beschwitz, B., D. B. Keim, and M. Massa, 2013, "Media-Driven High Frequency Trading: Evidence from News Analytics," *SSRN Working Paper Series*, pp. 1–46.

6 Tables and Figures

Table 1: Characteristics and Correlations of News Sentiment Variables

Panel A: Mean and standard deviation for various sentiment measures

Variable	Mean	Standard Deviation
Harvard Net Sentiment	2.4%	3.6%
Loughran-McDonald Net Sentiment	-2.5%	3.5%
Thomson Reuters Net Sentiment	2.4%	39.0%
Harvard Negative Sentiment	2.9%	2.6%
Harvard Positive Sentiment	5.3%	4.1%
Loughran-McDonald Negative Sentiment	2.9%	3.3%
Loughran-McDonald Positive Sentiment	0.5%	0.8%
Thomson Reuters Negative Sentiment	27.5%	24.6%
Thomson Reuters Positive Sentiment	29.9%	21.7%

Panel B: Correlation among net sentiment measures

	Harvard Net Sentiment	Loughran-McDonald Net Sentiment	Thomson Reuters Net Sentiment
Harvard Net Sentiment	1		
Loughran-McDonald Net Sentiment	-0.224	1	
Thomson Reuters Net Sentiment	0.151	0.249	1

Panel C: Correlation among sentiment measures

	Harvard Negative Sentiment	Harvard Positive Sentiment	Loughran-McDonald Negative Sentiment	Loughran-McDonald Positive Sentiment	Thomson Reuters Negative Sentiment	Thomson Reuters Positive Sentiment
Harvard Negative	1.00					
Harvard Positive	0.51	1.00				
Loughran-McDonald Negative	0.50	0.53	1.00			
Loughran-McDonald Positive	-0.11	-0.06	-0.23	1.00		
Thomson Reuters Negative	0.09	-0.12	0.01	0.10	1.00	
Thomson Reuters Positive	-0.21	-0.10	-0.44	0.27	-0.42	1.00

Notes: This table shows the average net firm sentiment (positive minus negative), positive sentiment and negative sentiment for 900,754 articles using the Harvard psychosocial dictionary, the Loughran-McDonald financial sentiment dictionary, and the Thompson-Reuters sentiment engine.

Table 2: Weekly Summary Statistics by Market Capitalization

Decile	Log Market Cap	News Stories Per Week	Proportion of firms without news	Return	Return w/news	Return w/o news	Difference	*t-statistics*
Smallest 1	9.42	0.16	0.95	-0.14%	2.00%	-0.24%	2.24%	3.47
2	10.56	0.59	0.94	-0.12%	1.51%	-0.23%	1.75%	3.59
3	11.22	1.33	0.92	-0.14%	-0.02%	-0.15%	0.12%	0.40
4	11.80	2.49	0.89	-0.06%	-0.04%	-0.06%	0.01%	0.04
5	12.35	4.01	0.87	-0.05%	-0.28%	-0.01%	-0.26%	-0.86
6	12.88	5.81	0.85	-0.01%	-0.26%	0.03%	-0.29%	-0.99
7	13.43	7.63	0.82	0.04%	-0.07%	0.06%	-0.13%	-0.48
8	14.04	10.19	0.76	0.05%	0.08%	0.04%	0.04%	0.15
9	14.83	13.30	0.66	0.09%	0.12%	0.07%	0.05%	0.25
Biggest 10	16.48	22.42	0.34	0.06%	0.07%	0.04%	0.04%	0.18

Notes: This table presents weekly statistics for decile portfolios grouped on market capitalization over calendar years 2003-2010 (417 weeks). We divide the firms into deciles based on the market capitalization at the beginning of the month. Each week we note the number of news articles with relevance of at least 0.35, proportion of firms without news, average return, average return for firms with news, average return for firms without news, difference of return between firms with and without news and the t statistics for the difference.

Table 3: Return from a Long-Short Portfolio Based on News in Lagged Days

Lagged Day	Harvard Quintiles		LM Quintiles		TR Quintiles	
-10	Mean	t	Mean	t	Mean	t
-9	0.01%	0.5	0.06%	3.0	0.09%	6.0
-8	0.01%	0.3	0.05%	2.7	0.07%	4.5
-7	0.05%	2.9	0.08%	4.4	0.09%	5.9
-6	0.03%	1.8	0.07%	3.8	0.10%	6.3
-5	0.05%	2.7	0.09%	5.1	0.12%	7.4
-4	0.03%	1.9	0.08%	4.1	0.08%	4.7
-3	0.09%	4.3	0.06%	3.0	0.12%	6.9
-2	0.04%	1.8	0.10%	5.0	0.18%	10.8
-1	0.18%	8.0	0.25%	10.0	0.50%	22.4
0	0.61%	19.4	1.21%	35.9	1.99%	63.9
1	-0.03%	-1.4	0.09%	4.1	0.17%	9.8
2	-0.01%	-0.7	-0.00%	0.0	0.04%	2.5
3	-0.01%	-0.4	0.02%	0.9	0.02%	1.2
4	0.00%	0.0	0.04%	2.0	0.04%	2.5
5	-0.02%	-0.9	0.02%	1.1	0.03%	1.6
6	0.01%	0.6	-0.01%	-0.6	0.06%	0.4
7	0.02%	0.9	0.02%	0.8	0.02%	1.1
8	-0.00%	-0.2	-0.00%	-0.2	0.01%	0.9
9	-0.00%	-0.3	-0.03%	-1.5	-0.02%	-1.2
10	0.01%	0.7	0.01%	0.6	-0.00%	0.0

Notes: We sort all stocks on a day based on the news sentiment from a lagged day and take a long position in the highest quintile (positive news stocks) and a short position in the lowest quintile (negative news stocks). This table shows the average daily return on long-short portfolio from sentiment scores using the Harvard psychosocial dictionary, the Loughran McDonald financial sentiment dictionary, and the Thomson-Reuters sentiment engine.

Table 4: Weekly Returns from Long-Short Portfolio Based on News in Lagged Weeks.

(a) Panel A: Weekly Returns from Long-Short Portfolio for all stocks with news

Lagged Week	Harvard	t	LM	t	Reuters	t
0	1.21%	14.5	2.20%	20.4	3.75%	37.0
1	0.12%	1.7	0.33%	4.3	0.32%	3.9
2	0.06%	0.9	0.07%	1.0	0.20%	2.6
3	0.09%	1.4	0.13%	1.8	0.26%	3.6
4	-0.07%	-1.1	0.07%	1.0	0.10%	1.4
5	-0.05%	-0.8	0.09%	1.4	0.19%	2.6
6	0.05%	0.8	0.08%	1.1	0.14%	1.9
7	-0.03%	-0.4	0.03%	0.4	0.11%	1.5
8	-0.02%	-0.3	0.08%	1.2	0.08%	1.2
9	0.06%	1.0	0.04%	0.7	0.12%	1.6
10	-0.05%	-0.8	-0.05%	-0.7	0.21%	2.8
11	0.07%	1.1	0.04%	0.7	0.20%	2.9
12	0.05%	0.8	0.01%	0.2	0.01%	0.2
13	0.00%	0.0	0.03%	0.5	0.21%	2.6

(b) Panel B: Weekly Returns from TR score based Long-Short Portfolio for stocks with one day of news in the week 0.

Lagged Week	Return	t statistics
0	3.37%	41.08
1	0.19%	3.13
2	0.11%	1.98
3	0.15%	2.92
4	0.08%	1.29
5	0.09%	1.67
6	0.09%	1.56
7	0.12%	2.09
8	0.06%	1.13
9	0.08%	1.33
10	0.11%	2.05
11	0.11%	2.06
12	-0.05%	-0.99
13	0.10%	1.69

(c) Panel C: Weekly Returns from TR score based Long-Short Portfolio for stocks with more than one day of news in the week 0.

Lagged Week	return	t statistics
0	4.22%	24.46
1	0.48%	3.94
2	0.21%	1.79
3	0.21%	1.86
4	-0.04%	-0.36
5	0.20%	1.85
6	0.10%	1.00
7	0.01%	0.14
8	0.13%	1.24
9	0.21%	1.88
10	0.13%	1.17
11	0.06%	0.64
12	0.07%	0.69
13	0.25%	2.32

Notes: We sort all stocks in a week based on the news sentiment from a lagged week and take a long position in the highest quintile (positive news stocks) and a short position in the lowest quintile (negative news stocks). Panel A shows the average weekly return on long-short portfolio using sentiment scores using the Harvard psychosocial dictionary (Harvard), the Loughran McDonald (LM) financial sentiment dictionary, and the Thomson Reuters (Reuters) sentiment engine. Panel B shows the return from a long-short portfolio which is constrained to stocks which had only one day of news in the week 0. Panel C shows the return from a long-short portfolio constrained to stocks with more than one day of news in the week 0.

Table 5: Cross-Sectional Regressions of Weekly Returns on Lagged Sentiment

(a) Average Cross-Sectional Regression Coefficient on Harvard Sentiment Variables

Lagged Week	α	t-value	News Effect	t-value	Positive Sentiment	t-value	Negative Sentiment	t-value
0	-0.0005	-0.3	-0.0006	-1.0	0.1157	15.4	-0.1939	-17.6
1	-0.0005	-0.3	0.0003	0.7	0.0088	1.6	-0.0153	-1.8
2	-0.0006	-0.4	0.0004	0.9	0.0080	1.7	-0.0052	-0.6
3	-0.0006	-0.4	0.0002	0.4	0.0092	1.9	0.0023	0.3
4	-0.0005	-0.3	0.0004	0.9	0.0053	1.1	0.0037	0.4
5	-0.0004	-0.3	0.0003	0.7	-0.0013	-0.2	0.0044	0.5
6	-0.0003	-0.2	0.0005	1.0	0.0028	0.6	-0.0013	-0.2
7	-0.0003	-0.2	0.0007	1.7	-0.0012	-0.2	0.0049	0.6
8	-0.0003	-0.2	0.0009	1.9	-0.0045	-0.8	-0.0030	-0.3
9	-0.0002	-0.2	0.0004	0.9	0.0042	0.8	-0.0036	-0.4
10	-0.0001	0.0	0.0002	0.5	0.0021	0.4	0.0079	1.0
11	-0.0004	-0.3	0.0000	0.0	0.0089	1.8	0.0026	0.3
12	-0.0004	-0.3	0.0006	1.4	0.0080	1.6	0.0034	0.4
13	-0.0002	-0.2	0.0009	1.9	-0.0013	-0.2	-0.0194	-2.3
					0.1646		-0.2126	

(b) Average Cross-Sectional Regression Coefficient on Loughran-McDonald Sentiment Variables

Lagged Week	α	t-value	News Effect	t-value	Positive Sentiment	t-value	Negative Sentiment	t-value
0			0.0030	7.0	0.3356	16.2	-0.1970	-14.9
1			0.0009	2.5	0.0325	2.2	-0.0278	-3.5
2			0.0007	1.9	-0.0074	-0.5	-0.0014	-0.2
3			0.0009	2.4	0.0006	0.0	-0.0109	-1.5
4			0.0008	2.2	-0.0131	-0.8	0.0032	0.4
5			0.0006	1.4	-0.0169	-1.2	-0.0003	0.0
6			0.0006	1.5	0.0042	0.2	0.0009	0.1
7			0.0010	2.5	-0.0298	-2.0	-0.0018	-0.3
8			0.0006	1.5	0.0015	0.1	0.0028	0.4
9			0.0007	1.6	-0.0167	-1.1	-0.0027	-0.4
10			0.0005	1.2	-0.0270	-1.6	0.0007	0.1
11			0.0005	1.2	-0.0146	-0.9	0.0022	0.3
12			0.0008	2.3	0.0043	0.3	0.0055	0.8
13			0.0005	1.3	-0.0228	-1.4	-0.0040	-0.6
					0.2305		-0.2307	

(c) Average Cross-Sectional Regression Coefficient on Thomson Reuters Sentiment Variables

Lagged Week	α	t-value	News Effect	t-value	Positive Sentiment	t-value	Negative Sentiment	t-value
0			-0.0003	-0.4	0.0304	27.6	-0.0329	-23.5
1			0.0008	1.4	0.0017	**2.0**	-0.0037	-3.7
2			0.0008	1.5	0.0008	0.9	-0.0018	-1.9
3			0.0012	2.2	0.0004	0.5	-0.0032	-3.5
4			0.0019	3.4	-0.0016	-1.9	-0.0027	-2.7
5			0.0012	2.2	-0.0003	-0.4	-0.0026	-2.9
6			0.0009	1.5	-0.0002	-0.2	-0.0018	-1.9
7			0.0010	1.8	0.0002	0.2	-0.0015	-1.6
8			0.0012	2.1	-0.0005	-0.7	-0.0018	-1.9
9			0.0004	0.7	0.0007	0.9	-0.0011	-1.1
10			0.0011	2.0	-0.0005	-0.7	-0.0026	-2.8
11			0.0003	0.6	0.0009	1.2	-0.0011	-1.3
12			0.0013	2.4	-0.0005	-0.7	-0.0009	-1.0
13			0.0004	0.8	0.0008	1.0	-0.0017	-1.8
					0.0321		-0.0593	

Notes: We use the cross-sectional regression technique of Fama and MacBeth (1973). For a given lag k ranging from 0 to 13, we regress stock returns on sentiment ratings

$$r_i(t) = \alpha_k(t) + \gamma_k(t) * 1_{If_{news}}(t-k) + \beta_k(t) * Positive_i(t-k) + \delta_k(t) * Negative_i(t-k) + \epsilon_i(t) \quad (2)$$

where $r_i(t)$ is the return on stock i in week t, $1_{If_{news}}(t-k)$ is a dummy variable for firms with news over the lag k, and $Positive_i(t-k)$, and $Negative_i(t-k)$ are the evaluation of positive and negative sentiment in news article published in the lagged week k. We report the time series average of $\gamma_k(t)$ in the 'News Effect' column and report the time series averages of $\beta_k(t)$ and $\delta_k(t)$ in the 'Positive Sentiment' and 'Negative Sentiment' columns.

Figure 1: Cumulative Daily News and Post-News Long-Short Quintile Returns

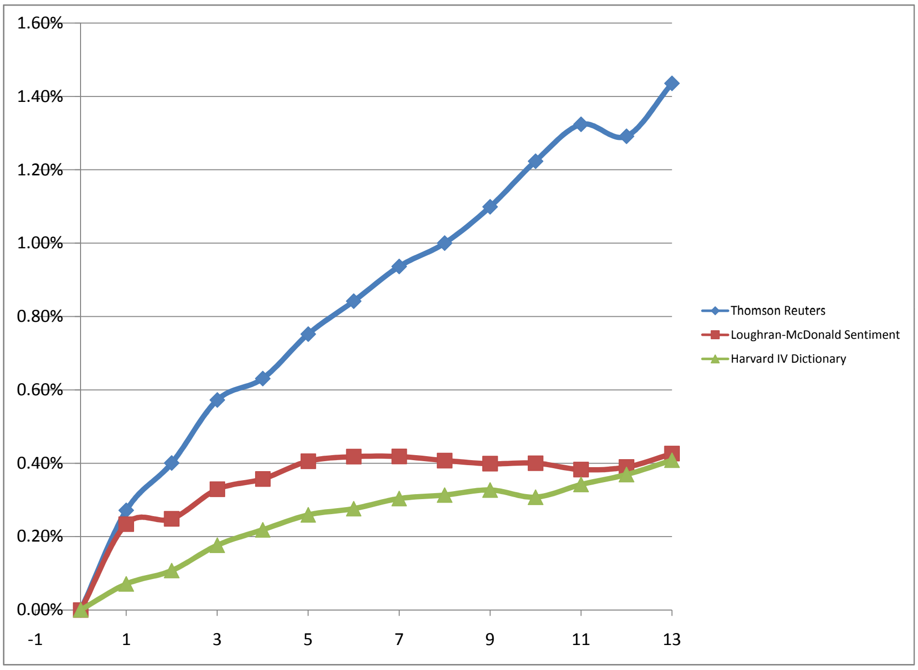
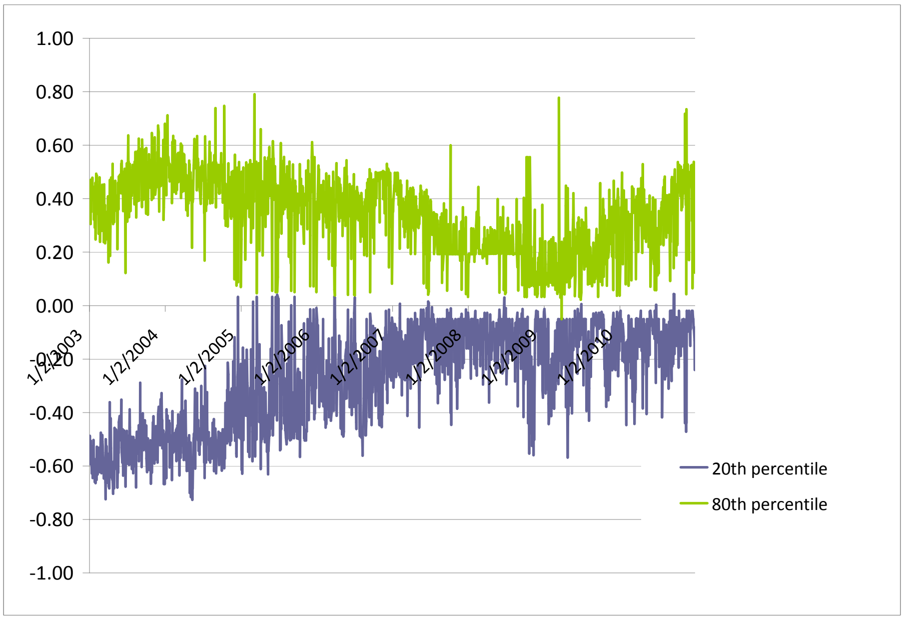
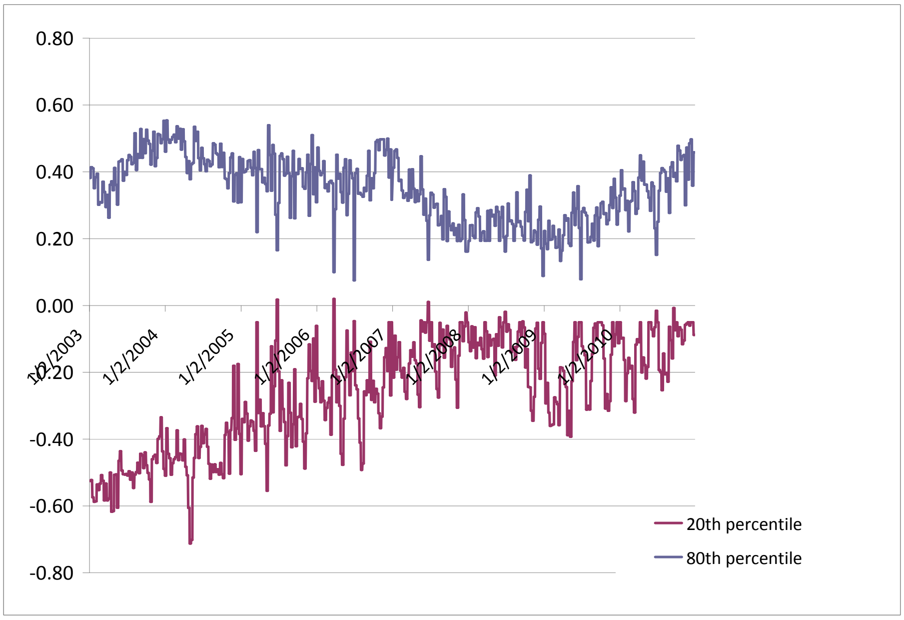


Figure 2: Cumulative Weekly Post-News Long-Short Quintile Returns



(a) Daily Fractile Levels for Thomson-Reuters News Sentiment



(b) Weekly Fractile Levels for Thomson-Reuters News Sentiment

Figure 3: Daily and Weekly Fractile Levels for Thomson-Reuters News Sentiment

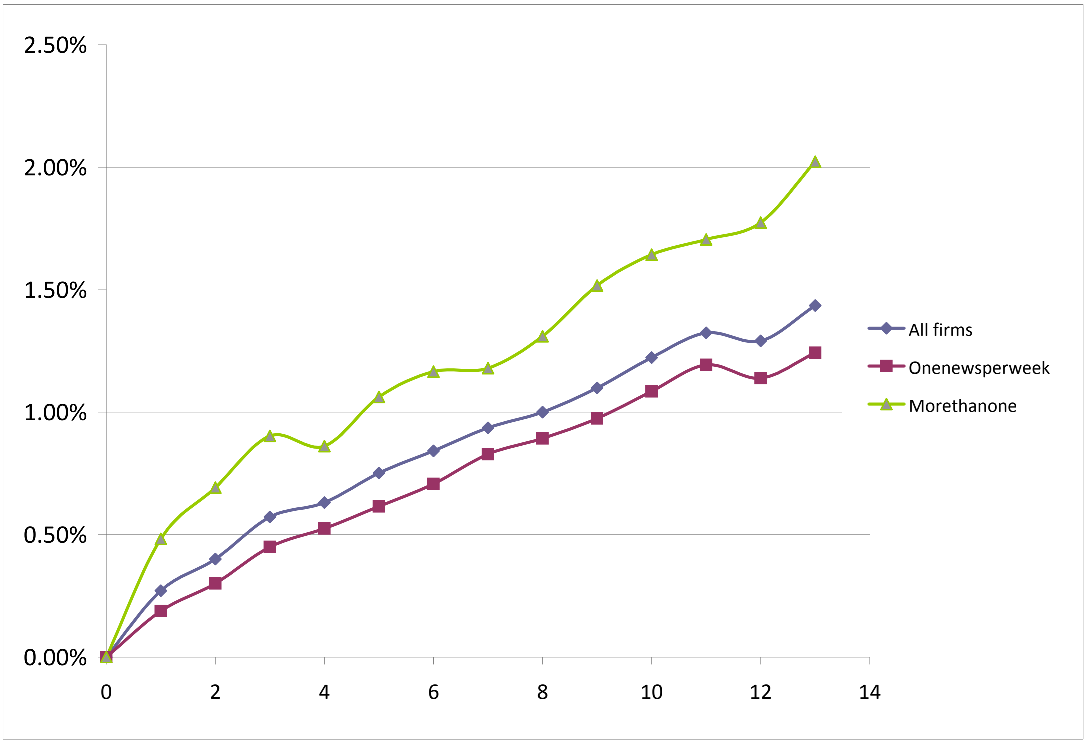
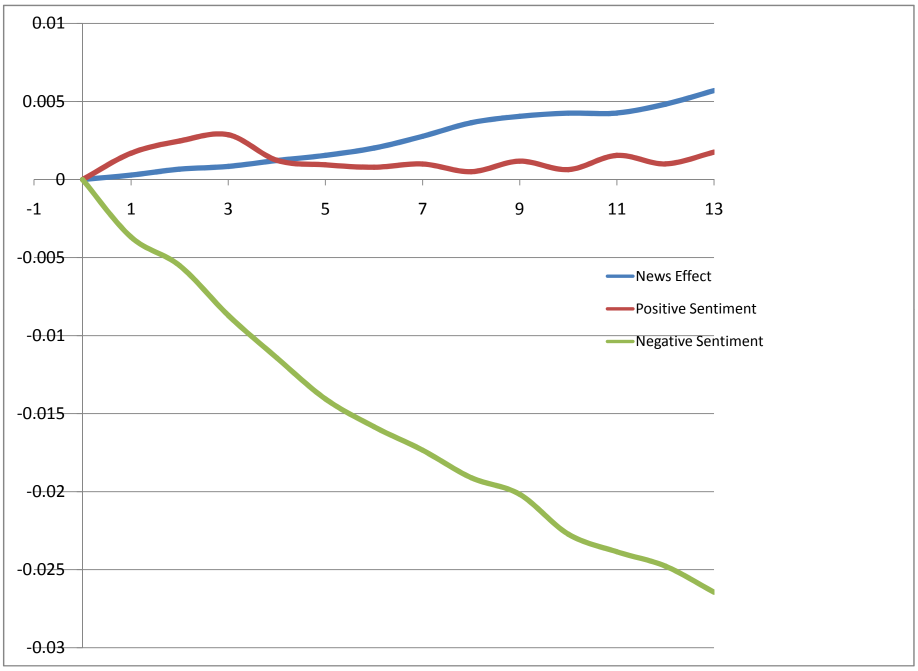


Figure 4: Cumulative Weekly Post-News Long-Short Quintile Returns

Figure 5: Cumulative Weekly Average Cross-Sectional Regression Coefficients



The figure plots the cumulative coefficients from Table 5 for horizons ranging from one week to thirteen weeks. The table reports time series average from the following regression. For a given lag k ranging from 0 to 13, we regress stock returns on sentiment ratings

$$r_i(t) = \alpha_k(t) + \gamma_k(t) * 1_{Ifnews}(t-k) + \beta_k(t) * Positive_i(t-k) + \delta_k(t) * Negative_i(t-k) + \epsilon_i(t)$$

where $r_i(t)$ is the return on stock i in week t, $1_{Ifnews}(t-k)$ is a dummy variable for firms with news over the lag k, and $Positive_i(t-k)$, and $Negative_i(t-k)$ are the evaluation of positive and negative sentiment in news article published in the lagged week k.