WHICH NEWS MOVES STOCK PRICES? A TEXTUAL ANALYSIS[1]

Jacob Boudoukh[2], Ronen Feldman[3], Shimon Kogan[4] and Matthew Richardson[5]

[PRELIMINARY]

Abstract:

A basic tenet of financial economics is that asset prices change in response to unexpected fundamental information. Since Roll's (1988) provocative presidential address that showed little relation between stock prices and news, however, the finance literature has had limited success reversing this finding. This paper revisits this topic in a novel way. Using advancements in the area of textual analysis, we are better able to identify relevant news, both by type and by tone. Once news is correctly identified in this manner, there is considerably more evidence of a strong relationship between stock price changes and information. For example, market model R^2s are no longer the same on news versus no news days (i.e., Roll's (1988) infamous result), but now are 15% versus 34%; variance ratios of returns on identified news versus no news days are 36% higher versus only 2% for unidentified news versus no news; and, conditional on extreme moves, stock price reversals on the order of 80 basis points occur on no news and unidentified news days, while identified news days show an opposite effect, namely a strong degree of continuation. A number of these results are strengthened further when the tone of the news is taken into account by measuring the positive/negative sentiment of the news story.

[1] Corresponding author: Shimon Kogan, GSB 5.159, McCombs School of Business, University of Texas at Austin, 1 University Station, B6600, Austin, TX 78712, Tel: +1 (512) 232-6839, email shimon.kogan@austin.utexas.edu. We would like to thank John Griffin and seminar participants at the University of Texas, Austin, and Stern NYU for their comments and suggestions.
[2] Arison School of Business, IDC, Israel
[3] School of Business Administration, The Hebrew University
[4] McCombs School of Business, University of Texas at Austin
[5] Stern School of Business, New York University and NBER

I. Introduction

A basic tenet of financial economics is that asset prices change in response to unexpected fundamental information. Early work, primarily though event studies, seemed to confirm this hypothesis. (See, for example, Ball and Brown (1968) on earning announcements, Fama, Fisher, Jensen and Roll (1969) on stock splits, Mandelker (1974) on mergers, Aharony and Swary (1980) on dividend changes, and Asquith and Mullins (1986) on common stock issuance, among many others.) However, since Roll's (1988) provocative presidential address that showed little relation between stock prices and news (used as a proxy for information), the finance literature has had limited success at showing a strong relationship between prices and news, e.g., also see Shiller (1981), Cutler, Poterba and Summers (1989), Campbell (1991), Berry and Howe (1994), Mitchell and Mulherin (1994), and Tetlock (2007), to name a few. The basic conclusion from this literature is that stock price movements are largely described by irrational noise trading or through the revelation of private information through trading.

In this paper, we posit an alternative explanation, namely that the finance literature has simply been doing a poor job of identifying true and relevant news. In particular, common news sources for companies such as those in the Wall Street Journal stories and Dow Jones News Service, et cetera, contain many stories which are not relevant for information about company fundamentals. The problem of course is for the researcher to be able to parse through which news stories are relevant and which are not. Given that there are hundreds of thousands, possibly millions, of news stories to work through, this presents a massive computational problem for the researcher. Fortunately, advances in the area of textual analysis allow for better identification of relevant news, both by type and tone. This paper employs one such approach based on an information extraction platform (Feldman, Rosenfeld, Bar-Haim and Fresko (2011), denote Feldman at al. (2011)).

There is a growing literature in finance that uses textual analysis to try and convert qualitative information contained in news stories and corporate announcements into a quantifiable measure by analyzing the positive or negative tone of the information. One of the earliest papers is Tetlock (2007) who employs the General Inquirer, a well-known textual analysis program, alongside the Harvard-IV-4 dictionary to calculate the fraction of negative words in the *Abreast of the Market* Wall Street Journal column. Numerous papers have produced similar analyses to measure a document's tone in a variety of financial and accounting contexts, including Davis, Piger, and Sedor (2006), Engelberg (2008), Tetlock, Saar-Tsechansky and Macskassy (2008), Demers and Vega (2010), Feldman, Govindaraj, Livnat and Segal (2010), and Loughran and McDonald (2011), among others. While all these papers support the idea that news, transformed into a sentiment measure, have important information for stock prices, none represent a significant shift in thinking about the overall relation between stock prices and information. Part of the reason is that, other than refinements of the Harvard-IV-dictionary for financial applications (e.g., Engelberg (2008) and Loughran and McDonald (2011)), the textual analysis methodology is similar.[6]

The aforementioned textual analysis methodology (Feldman et al. (2011)) employed in this paper is quite different. It combines not only a dictionary-based sentiment measure as in Tetlock (2007) and Loughran and McDonald (2011), but also an analysis of phrase-level patterns to further break down the tone of the article and a methodology for identifying relevant events for companies (broken down into 14 categories and 56 subcategories). While the methodology is for the most part based on sets of rules (as opposed to say machine learning),[7] the implementation employs the commonly used technique of running and refining these rules on a subset of training articles. This procedure greatly improves the accuracy. In terms of relating stock prices to news, the methodology provides a number of advantages over existing approaches. In particular, over the sample period 2000-2009 for all S&P500 companies, the Dow Jones Newswire produces over 475,000 stories, only 30% of which we identify as relevant events. As discussed shortly, this breakdown into identified and unidentified news makes a massive difference in terms of our understanding of stock price changes and news. Moreover, employing a more sophisticated textual analysis methodology than one based on a simple count of positive versus negative words (e.g., as in Tetlock (2007)) further improves the results. In other words, when we can identify the

[6] Some exceptions include Li (2010), Hanley and Hoberg (2011), and Grob-Klubmann and Hautsch (2011) who all use some type of machine learning-based application.
[7] Some parts of the implementation, such as locating names of companies and individuals, employ machine-learning technology, that is, the use of statistical patterns to infer context.

news, and more accurately evaluate its tone, there is considerably more evidence of a strong relationship between stock price changes and information.

This paper documents several new results. First, and foremost, using the aforementioned methodology that allows us to automatically and objectively classify articles into topics (such as analyst recommendations, financial information, acquisitions and mergers, etc.), we compare days with no-news, unidentified news, and identified news on several dimensions. In particular, we show that stock-level volatility is similar on no-news days and unidentified news days, consistent with the idea that the intensity and importance of information arrival is the same across these days. In contrast, on identified news days, the volatility of stock prices is 150% higher. This evidence is provided further support by noting that identified news days are 25-30% more likely to be associated with extreme returns (defined by the bottom and top 10% of the return distribution) while unidentified and no news days are slightly more likely to be associated with moderate day returns (in the middle 30-70% range of the returns distribution). A major finding is that when we revisit Roll's (1988) R^2 methodology and estimate the R^2 from a market model regression for all days and for unidentified news days, consistent with his results, R^2 levels are the same for all days and for unidentified news days. However, when we estimate the same model over just identified news days, the R^2 drops dramatically from an overall average of 28% to 19%, the precise result that Roll (1988) was originally looking for in his work. Even further support is provided when we additionally parse out identified news days into those with new versus stale news, leading to 77% higher stock price volatility and further R^2 drop of 21%.

Second, beyond the parsing of news into identified events and unidentified news, the methodology provides a measure of article tone (that is, positive versus negative) that builds on Tetlock (2007) and others. As mentioned above, we perform both an analysis of phrase-level patterns (e.g., by narrowing down to the relevant body of text, taking into account phrases and negation, etc.) and employ a dictionary of positive and negative words more appropriate for a financial context. Using this more advanced methodology, in contrast to a simple word count (e.g., as used by Tetlock (2007)), we show that our measure of tone can substantially increase R^2 on identified news days, but not on unidentified news days, again consistent with the idea that identified news days contain price-relevant information. Another finding is that tone variation across topics and within topics is consistent with one's intuition. For example, deals and partnership announcements tend to be very positive while legal announcements tend to be negative. Analyst recommendations and financial information, on average, tend to be more neutral, but tend to have greater variation within the topic. Moreover, some of these topics are much more likely to appear on extreme return days (e.g., analyst recommendations, financials) while others are not (e.g., partnership). This suggests that different topics may have different price impact. Finally, the results are generally consistent with a positive association between daily returns and daily tone, with this relationship being more pronounced using the methodology presented here than of the more standard simple word count.

Third, the above discussion contemporaneously relates relevant news to stock price changes. An interesting issue is whether the differentiation between identified and unidentified news has forecast power for stock price changes. There is now a long literature, motivated through work in behavioral finance and limits of arbitrage, that stock prices tend to underreact or overreact to news, depending on the circumstances (see, for example, Hirshleifer (2000), Chan (2003), Vega (2006), Gutierrez and Kelley (2008), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Tetlock (2010)). This paper documents an interesting result in the context of the breakdown of Dow Jones news into identified and unidentified news. Specifically, conditional on extreme moves, stock price reversals on the order of 80 basis points occur on no news and unidentified news days, while identified news days show an opposite effect, namely a small degree of continuation. That news days tend to be associated with future continuation patters while no news days see reversals is consistent with (1) our methodology correctly parsing out relevant news, and (2) a natural partition between underreaction and overreaction predictions in a behavioral context. As an additional test, we perform an out-of-sample exercise based on a simple portfolio strategy. The resulting gross Sharpe ratio of 1.4 illustrates the strength of these results.

While our paper falls into the area of the literature that focuses on using textual analysis to address the question of how prices are related to information, the two most closely related papers to ours, Griffin, Hirschey and Kelly (2011) and Engle, Hansen and Lunde (2011), actually lie outside this textual analysis area. Griffin, Hirschey and Kelly (2011) cross-check global news stories against earnings announcements to try and uncover relevant events. Engle, Hansen and Lunde (2011) utilize the Dow Jones Intelligent Indexing product to match news and event types for a small set of (albeit large) firms. While the focus of each of these papers is different (e.g., Griffin, Hischey and Kelly (2011) stress cross-country differences and Engle, Hansen and Lunde (2011) emphasizing the dynamics of volatility based on information arrival), both papers provide some evidence that better information processing by researchers will lead to higher R^2s between prices and news.

This paper is organized as follows. Section II describes the data employed throughout the study. Of special interest, we describe in detail the textual analysis methodology for inferring content and tone from news stories. Section III provides the main results of the paper, showing a strong relationship between prices and news, once the news is appropriately identified. In section IV, we reexamine a number of results related to the existing literature measuring the relationship between stock sentiment and stock returns. Section V discusses and analyzes the forecasting power of the textual analysis methodology for future stock prices, focusing on continuations and reversals after large stock price moves. Section VI concludes.

II. Data Description and Textual Analysis Methodology

A. Textual Analysis

With the large increase in the amount of daily news content on companies over the past decade, it should be no surprise that the finance literature has turned to textual analysis as one way to understand how information both arrives to the marketplace and relates to stock prices of the relevant companies. Pre mainstream finance, early work centered on document-level sentiment classification of news articles by employing pre-defined sentiment lexicons.[8] The earliest paper in finance that explores textual analysis is Antweiler and Frank (2005) who employ language algorithms to analyze internet stock message

[8] See, for example, Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, and Allan (2000), Das and Chen (2007) and Devitt and Ahmad (2007), among others. Feldman and Sanger (2006) provide an overview.

boards posted on "Yahoo Finance". Much of the finance literature, however, has focused on word counts based on dictionary-defined positive versus negative words.

For example, one of the best known papers is Tetlock (2007). Tetlock (2007) employs the General Inquirer, a well-known textual analysis program, alongside the Harvard-IV-4 dictionary to calculate the fraction of negative words in the *Abreast of the Market* Wall Street Journal column. A plethora of papers, post Tetlock (2007), apply a similar methodology to measure the positive versus negative tone of news across a wide variety of finance and accounting applications.[9] Loughran and McDonald (2011), in particular, is interesting because they refine the Harvard-IV-4 dictionary to more finance-centric definitions of positive and negative words.[10]

More recently, an alternative approach to textual analysis in finance and accounting has been offered by Li (2010), Hanley and Hoberg (2011), and Grob-Klubmann and Hautsch (2011). These authors employ machine learning-based applications to decipher the tone and therefore the sentiment of news articles. The basic approach of machine learning is not to rely on written rules per se, but instead allow the computer to apply statistical methods to the documents in question. In particular, supervised machine learning uses a set of training documents (that are already classified into a set of predefined categories) to generate a statistical model that can then be used to classify any number of new unclassified documents. The features that represent each document are typically the words that are inside the document (bag of words approach).[11] While machine learning has generally come to dominate rules-based classification approaches (that rely solely on human-generated rules), there are disadvantages, especially to the extent that machine learning classifies documents in a non transparent fashion that can lead to greater misspecification.

In this paper, in contrast, classification is not used at all. Instead, a rule based information extraction approach is employed, appealing to recent advances in the area of textual analysis

[9] See, for example, Davis, Piger, and Sedor (2006), Engelberg (2008), Tetlock, Saar-Tsechansky and Macskassy (2008), Kothari, Li and Short (2009), Demers and Vega (2010), Feldman, Govindaraj, Livnat and Segal (2010), and Loughran and McDonald (2011), among others.

[10] For a description and list of the relevant words, see http://nd.edu/~mcdonald/Word_Lists.html.

[11] See Manning and Schutze (1999) for a detailed description and analysis of machine learning methods.

(Feldman at al. (2011)) . That is, we extract event instances out of the text based on a set of predefined rules. For instance, when we extract an instance of an *Acquisition* event, we find who is the acquirer, who is the acquiree, optionally what was the amount of money paid for the acquisition, and so forth. Feldman et al. (2011) employ a proprietary information extraction platform specific to financial companies, which they denote *The Stock Sonar* (TSS), and which is available on commercial platforms like Dow Jones. This textual analysis methodology differs from current rules-based applications in finance in three important ways.

First, TSS also adheres to a dictionary-based sentiment analysis. In particular, the method uses as a starting point the dictionaries used by Tetlock (2007) and Loughran and McDonald (2011), but then augments by adding and subtracting from these dictionaries. Beyond the usual suspects of positive and negative words, a particular weight is placed on sentiment modifiers such as “highly”, “incredible”, “huge”, et cetera versus lower emphasis modifiers such as “mostly” and “quite” versus opposite modifiers such as “far from”. These words were adjusted to the domain of financial news by adding and removing many terms, depending on the content of thousands of news articles. Specifically, for developing these lexicons and rules (to be discussed in further detail below), a benchmark consisting of thousands of news articles was manually tagged. The benchmark was divided into a training set (providing examples) and a test set (kept blind and used for evaluating the progress of the methodology). The rulebook was run repeatedly on the system on thousands of articles, each time revised and iterated upon until the precision was satisfactory (e.g., >90%).

Second, this same approach was used to create a set of rules to capture phrase-level sentiments. Current systems employed in finance so far have operated for the most part at the word level, but compositional expressions are known to be very important in textual analysis. For example, one of the best known illustrations involve double negatives such as “reducing losses” which of course has a positive meaning, yet would likely yield a negative word count in most schemes. Other examples include words like “despite” which tend to connect both positive and negative information. A large number of expressions of this sort are considered jointly with the word dictionary to help better uncover the sentiment of the article.

Third, and most important, TSS sorts through the document and parses out the meaning of the document in the context of possible events relevant to companies, such as new product launches, lawsuits, analyst coverage, financial news, mergers, et cetera. The initial list of events were chosen to match commercial providers such as CapitalIQ but were augmented by events likely to impact stock prices. This process led to a total of 14 event categories and 56 subcategories within events. For example, the events fall into one of the following categories: *Analyst Recommendations*, *Financial*, *Financial Pattern*, *Acquisition*, *Deals*, *Employment*, *Product*, *Partnerships*, *Inside Purchase*, *Facilities*, *Legal*, *Award*, *Stock Price Change* and *Stock Price Change Pattern*. Consider the *Analyst Recommendation* category.[12] In terms of subcategories, it contains nine subcategories, including analyst expectation, analyst opinion, analyst rating, analyst recommendation, credit - debt rating, fundamental analysis, price target, etc.[13]

Because events are complex objects to capture in the context of textual analysis of documents, considerable effort was applied to write rules that can take any news story and then link the name of a company to both the identified event and sentiment surrounding the event. For each company, TSS identifies the exact body of text within the document that refers to that company so that the sentiment calculations will be based just on words and phrase that are directly associated with that company. For example, one specific technique is to consider only words within a range of the mention of the main company in the document. Another is to avoid historical events cited in documents by capturing past versus present tenses. Like the document sentiment analysis, a training set of documents were used to refine the rulebook for events and then evaluated against a test set.

[12] In practice, the categories, defined in terms of *Pattern*, represent cases in which an event was identified but the reference entity was ambiguous.

[13] For a complete list of the categories and subcategories, see http://shimonkogan.tumblr.com.

B. Data Description and Summary

The primary dataset used in this paper consists of all documents that pass through the Dow Jones Newswire from January 1, 2000 to December 31, 2009. For computational reasons, we limit ourselves to the S&P500 companies at the time the news stories are released. Over the sample period, the dataset therefore includes at some time or another 795 companies. To avoid survivorship bias, we include in the analysis all stocks in the index as of the first trading day of each year. We obtain total daily returns from CRSP.

TSS methodology described in II.A processes each article separately and generates an output file in which each stock/article/date is represented as an observation. For each of these observations, TSS reports the total number of words in the article, the number of relevant words in the article, the event (and sub-event) identified, and the number of positive and negative features as identified by TSS. For the same set of articles we also count the number of positive and negative words using the Harvard Dictionary IV (see, for example, Tetlock (2007)). In terms of sentiment score, after parsing out only relevant sentences, and determining the appropriate context of words at the phrase-level, the sentiment score is analyzed through the standard method of summing up over positive and negative words, e.g., $S = \frac{P-N}{P+N+1}$, where P and N stand for the number of positive and negative words, respectively.

A key part of this paper is being able to differentiate between relevant news for companies (defined in our context as those related to specific firm events) as opposed to unidentified firm events. For each news story, therefore, our application of TSS produces a list of relevant events connected to this company and to this particular piece of news. It is possible that multiple events may be connected to a given story. In our analysis we ignore the *Stock Price Change* and *Stock Price Change Pattern* categories as these categories do not, on their own, represent fundamental news events. We also ignore *Award*, *Facilities*, and *Inside Purchase*, since these categories do not contain a sufficient number of observations. We are therefore left with eight main categories.

To be more precise, our goal is to analyze the difference in return patterns based on the type of information arrival. We therefore classify each day into one of three categories:

1. No news – observations without news coverage.
2. Unidentified news – observations for which none of the news coverage is identified.
3. Identified news – observations for which at least some of the news coverage is identified by at least one of the above events.

Moreover, we define "new" news versus "old" news by whether the news identifies the same event that had been identified in similar recent news stories of that company.[14] Specifically, a given event coverage is considered to be "new" if coverage of the same event type (and the same stock) is not identified during the previous five trading days.

Since our goal is to relate information arrival to stock returns, which are observed at the stock/day level, we rearrange the data to follow the same stock/day structure. To do that, we consolidate all events of the same type (for a given stock/date) into a single event by averaging their scores. The resulting dataset is structured such that for each stock (and date) we have a set of indicators denoting which events were covered for that stock/day, scores for each of the event types (when no-missing), and the daily score computed by adding the number of positive and negative features across all relevant articles. In order to ensure that the analysis does not suffer from a look-ahead bias, we use the article timestamp to match with the trading day. That is, we consider date *t* articles those that were released between 15:31 on date *t-1* and 15:30 on date *t*. Date *t* returns are computed using closing prices on dates *t-1* and *t*.

Table 1 provides an overview of the data. The first column in panel A reports the number of observations under each of the day classifications. First, we see that most days have no news coverage, i.e., 758,393 of the observations across 1,235,103 firm/day observations contain no news reported on the Dow Jones Newswire. Second, and most important, the majority of the days with news coverage, 334,990 of 476,710, do not have a single

[14] See Tetlock (2011) for a different procedure for parsing out new and stale news.

identified event. The different day types are spread across virtually all stocks (e.g., see second column), allowing us to look at differences in patterns across day types within a stock. The last three columns of the table provide an overview of the article and word counts across the different day types. We observe that identified news days are characterized by a larger number of articles and more relevant words per article compared with unidentified news days, i.e., 5.9 versus 2.5, and 412 versus 107, respectively. However, the difference in the average article length is less dramatic across the day types.

Panel B reports the average firm returns, market returns, and factor characteristics (size, book-to-market, and momentum) of observations across day types. Consistent with the prior literature, we find that firm size is correlated with media, even if this effect is small for our sample of S&P500 firms (quintile assignment of 4.50 for no news days vs. 4.72 for unidentified news days and 4.76 for unidentified news days). Importantly, return and factor characteristics are very similar for identified and unidentified news days. In unreported results, we considered a fourth category – days with both identified and unidentified news. The results are unaffected by merging this category into purely identified news days.

A key finding of the paper is that when we can identify news, they matter. As a first pass at the data, Table 2 provides a breakdown of news stories by the distribution of returns. In brief, the main result is that identified news are more likely than unidentified news to lie in the negative and positive tails of the return distribution. On the surface, this is consistent with rational models, which would suggest that information arrival should be associated with increases in volatility.

In particular, if news days proxy for information arrival, we should find that news arrival would be concentrated among days with large return movements, positive or negative. To relate news arrival intensity with returns, we assign daily returns into percentiles separately for each stock and year: bottom 10%, next 20%, middle 40%, next 20%, and top 10%. The columns in Table 2 group observations according to this split. The first three rows of the table show that extreme day returns are associated with somewhat larger number of articles (for each stock appearing in the news) and on these days, there is a larger total number of words used in the articles.

Next, we compare the observed intensity of different day types to the intensity predicted under the null that these distributions are independent. For example, the null would suggest that of the 758,393 no news days, 75.8 thousand would coincide with returns at the bottom 10%, 161.6 thousand would coincide with returns at the following 20%, and so forth. The results in rows five through fourteen report the difference between the observed intensity and the null in percentage terms.

Several observations are in order. First, we find that no news days are less concentrated among days with large price changes: -3.5% (-5.2%) for the bottom (top) 10% of days. This is consistent with the notion that news coverage proxies for information arrival. Interestingly though, we observe a very similar pattern for unidentified news days: -5.2% (-3.6%) for the bottom (top) 10% of days. Second, in sharp contrast to these results, we find that identified news days are 27% (31%) more likely to coincide with the bottom (top) 10% of return days. Thus, while we might expect under independence to have 14,172 identified news stories in the lower tail, we actually document 17,956 news stories. That is, identified news days, but not unidentified news days, are much more likely to be extreme return days.

Third, this last pattern is also observed when we examine the frequency of individual event types, one at a time. The bottom panel of Table 2 shows a U-shaped pattern suggesting that each of the event types is more likely to coincide with extreme return days compared with moderate return days. It should be noted that for some event types, the pattern is not symmetric. For example, "deals" are more likely to appear on extreme positive days, compared with extreme negative days. This is consistent with the intuition that deals would generally be regarded as a positive event for the firm. At the same time, "legal" event are more likely to coincide with extreme negative days compared with extreme positive days. The news categories with the greatest concentration of events in the tails – "Analyst Recommendations" and "Financial" – are not surprisingly dispersed in a much more symmetric way.

III. R^2

A seminal paper on the question of whether stock prices reflect fundamental information is Roll (1988). In that paper, Roll (1988) argues that once aggregate effects have been removed from a given firm, the finance paradigm would imply that the remaining variation of firm returns would be idiosyncratic to that firm. As a proxy for this firm specific information, Roll (1988) uses news stories generated in the financial press. His argument is that, on days without news, idiosyncratic information is low, and the R^2s from aggregate level regressions should be much higher. Roll (1988) finds little discernible difference. Thus, his conclusion is that it is difficult to understand the level of stock return variation. Working off this result, a number of other papers reach similar conclusions with respect to prices and news, in particular, Cutler, Poterba and Summers (1989), and Mitchell and Mulherin (1994).

The evidence that asset prices do not reflect seemingly relevant information is not just found with equity returns. For example, Roll (1984)'s finding that, in the frozen concentrated orange juice (FCOJ) futures market, weather surprises explain only a small amount of variability of futures returns has been a beacon for the behavioral finance and economics literature. Given that weather has theoretically the most important impact on FCOJ supply, and is the focus of the majority of news stories, Roll (1984) concludes, like in his 1988 paper, that there are large amounts of "inexplicable price volatility". In contrast, Boudoukh, Richardson, Shen and Whitelaw (2007) show that when the fundamental is identified, in this case temperatures close to or below freezing, there is a close relationship between prices and weather surprises. In this section, we make a similar argument to Boudoukh, Richardson, Shen and Whitelaw (2007). We parse out news stories into identified versus unidentified events and reevaluate Roll's (1988) finding and conclusion.

In a different context, and using a different methodology, Griffin, Hirschey and Kelly (2011) and Engle, Hansen and Lunde (2011) also provide evidence that price volatility can be partially explained by news. For example, by cross-checking global news stories against earnings announcements to try and uncover relevant events, Griffin, Hirschey and Kelly (2011) document better information extraction can lead to higher R^2s between prices and

news. Engle, Hansen and Lunde (2011) utilize the *Dow Jones Intelligent Indexing* product to match news and event types for a small set of (albeit large) firms, and show that the arrival of this public information has explanatory power for the dynamics of volatility.
The results of Table 2 suggest that our textual analysis methodology will have similar success at linking identified events to stock return variation.[15] Therefore, as a more formal look at the data, we study the link between news arrival and volatility by computing daily return variations on no news days, unidentified news days, and identified news days. Specifically, for each stock we compute the average of squared daily returns on these day types and then calculate the ratio of squared deviations on unidentified news days and no news days, and the ratio of squared deviations on identified news days and no news days (subtracting 1 from both ratios). If both unidentified and identified news days have no affect on stock volatility we should find that these ratios are distributed around zero.

Figure 1 depicts the distribution of ratios across the 671 stocks for which these ratios are available (out of 795), winsorized at 10.[16] As evident, the ratios are not distributed around zero for neither unidentified nor identified news days. However, the difference in distributions between unidentified and identified news days' ratios is clear: the variance ratio is much higher on identified news days compared with unidentified news days. On average, squared deviations are 30% higher on unidentified news days relative to no news days, while they are 187% higher on identified news days. These results clearly demonstrate that our day classification has power to distinguish between days on which price-relevant information arrives and days on which information may or may not arrive, but if it does, it is not price-relevant.

Table 3 compares daily percentage return squared variations sorted based on day type (no news, unidentified news, and identified news) and then by event types (*acquisition*, *analyst recommendations*, etc.). Consider the median squared returns across these classifications: 1.25 for no news days, somewhat higher, 1.28, on unidentified news days, and sharply

[15] Note that, while most researchers focus on Roll's (1988) R^2 result, Roll (1988) also provided evidence that kurtosis was higher on news versus no news days, a result similar to that provided in Table 2.
[16] We eliminate stocks for which we do not have at least twenty trading days of under each of the day categories.

higher, 1.70 on identified news days.[17] One of the useful applications of the TSS methodology is the ability to further parse out identified news into different types. A closer examination of variance patterns across event types suggest that some of the events coincide with particularly large identified news day variance – *analyst recommendations* (3.01) and *financial* (2.42). Also, consistent with the hypothesis that identified news represent price-relevant pieces of information, we find that among identified news days there is a substantial difference between the variance of old news days, with a median squared percentage return of 1.44, and new news days, with the corresponding statistics of 1.82.

Table 4 reports results for a reinvestigation of the aforementioned R^2 analysis of Roll (1988). Specifically, we estimate a one-factor pricing model separately for each firm and for each day classification: no news, unidentified news, and identified news.[18] We repeat the same analysis at the 2-digit SIC industry classification thereby imposing a single beta for all firms within a given industry and utilizing weighted least squared regressions.

The results in Panel A of Table 4 report the average and median R^2 across firms (columns 1 to 3) and industries (columns 4 to 6). Consider the median calculations. Under both specifications, the R^2s are similar on no news and unidentified news days (34% vs. 29%), consistent with Roll's puzzling results. However, R^2s are much lower on identified news day, i.e., 15%. The difference in R^2 between identified news and no-news days is striking and stands in sharp contrast to Roll's results. Roll's original conjecture, refuted by his 1988 work, was that the performance of a market model, as measured by R^2, should be much worse during days on which firm-specific information arrives compared with days when no such information arrives. Our results lend support to his conjecture since we are able to better proxy for such days using event identification.

As shown by Tables 2 and 3, and consistent with a priori intuition, not all events have the same informational impact on stock prices. It is worthwhile therefore to further explore the R^2s by breaking up news into event types. Panel B of Table 4 reports the market regression

[17] Note that this analysis pulls observations across stocks and in that way differs from the distribution results reported in Figure 1.
[18] We impose a minimum of 20 observations to estimate the regression.

of Roll (1988) conditional on each event type and whether the news is classified as new. The results show a large degree of variation across events. For example, acquisitions (11%), analyst recommendations (14%), financial (11%) and legal (14%) are lower than the 15% cited above for identified news days, and substantially lower than the 34% on no news days and 29% on unidentified news days. To the extent these categories can be further broken down, and the sentiment of each event incorporated, one would expect an even greater bifurcation of the R^2s between unidentified/no news days and further refined identified news days.

IV. Measuring Sentiment

One of the main applications of textual analysis in finance has been to link sentiment scores to both contemporaneous and future stock returns. The evidence is statistically significant albeit weak in magnitude. For example, Tetlock (2007) and Tetlock, Saar-Tsechansky and Macskassy (2008), show that negative word counts of news stories about a firm based on the Harvard IV dictionary have contemporaneous and forecast power for the firm's stock returns though the R^2s are low. Loughran and McDonald (2011) argue that for a finance context the Harvard dictionary is not appropriate and build a sentiment score using a more finance-centric dictionary. Their application focuses on creating a dictionary appropriate for understanding the sentiment contained in 10-K reports. For their 10-K application, sentiment scores based on word counts from this alternative dictionary generally provide a better fit.

In this section, we first extend the analysis of Section 3 on news versus no news R^2s to include sentiment scores. In the above analysis, we showed that identified news days are a good proxy for information arrival. Below, we show that the sentiment of these articles, i.e., the directional content of this information, has explanatory power for returns. As a preview, consider Table 4. Table 4 shows that market model regressions on news days have low R^2, that is, most of the variation of stock returns is idiosyncratic in nature. A reasonable hypothesis is that the R^2s should be increased if the idiosyncratic information is incorporated directly. We use as our proxy for this direct information the sentiment score, and we compare the score based on TSS and that using the Harvard IV-4 dictionary. These results are reported in Table 4.

To see the additional explanatory power of event-specific scores, consider the results at the bottom of Panel A of Table 4. The R^2s reported in the table derive from either firm level or industry regressions, while augmenting the one-factor market model with daily scores obtained from TSS or IV4. The industry level weighted least squared regressions augment the one-factor market model with event-level scores estimated at the 2-digit SIC industry level. That is, in the industry level regressions we assume that all firms within the industry have the same return response magnitude to a given event type but we allow this magnitude to vary across events and industries. Focusing on identified event days, we see that at the firm level, daily scores obtained from TSS increase R^2 from a median of 15% to 18% (while unimproved using IV-4 scores). The increase in R^2s at the industry level, when using even specific scores, is much more pronounced – R^2s increase from 15% to 20%, a 33% increase. Most important, these increases are attained only for identified news days. In contrast, for unidentified news days, there is almost no increase in R^2s when sentiment scores are taken into account. In other words, to link stock prices to information, it is necessary to measure both the news event and the tone (i.e., sentiment) of this news.

In order to investigate this further, Panel B of Table 4 reports R^2 from weighted least squared pooled industry regressions while separating observations by new news and event types. Even more striking results are reported in Panel B of the table. Compared with the performance of the baseline CAPM model, Table 4, Panel B shows that the TSS sentiment score, when allowing for event-specific scores, increases the explanatory power significantly. For example, median daily return R^2s exceed 30% for *analyst recommendations* (32%), *deals* (37%), *employment* (33%), *legal* (33%), *partnerships* (46%), and *product events* (41%). These results are especially impressive given the 15% R^2s without taking into account the tone of the stories.

In the remainder of this section, we take a step back and explore TSS sentiment scores in greater detail. Recall that for each day and event type (within the day) we compute a sentiment score using the number of positive and negative features identified by TSS. For comparison purposes, we also compute a score using the Harvard-IV-4 dictionary, similar to Tetlock (2007). We refer to these scores as "IV4". Table 5 provides a set of summary statistics with respect to sentiment scores.

The first column in the table reports the number of observations classified as unidentified and identified news days (first two rows), followed by the number of observations falling into each of the event types.[19] The set of columns under "TSS" report score statistics for each of the classifications. For example, of the 335 thousand unidentified news days, TSS is able to compute a sentiment score for only 121 thousand. In contrast, virtually all identified news days are matched with sentiment output from TSS. The remaining columns in the column block report the mean, percentiles (10%, 50%, 90%), and spread between the top and bottom 10% of observations within each category. The next block of columns, under "IV4", reports the same set of statistics using the IV-4 based dictionary. The last column in the table reports the correlation between the TSS and IV4 scores.

First, for virtually every category, the number of observations with available TSS scores is smaller than the number of observations with available IV4 scores available. This is consistent with the set of negative and positive words in the IV4 dictionary being generally larger than the set of positive and negative features in TSS. The average score for unidentified and identified news days is on average positive, demonstrating the tendency of media coverage to have a positive tone. This bias is similar in magnitude for TSS and IV4.

Second, TSS appears to produce more discerning sentiment scores compared with IV4. For both unidentified and identified days, the spread of TSS scores is much larger than the spread of IV4 scores; the difference between the top and bottom 10% of identified news days is 1.22 under TSS but only 0.59 under IV4.[20] This holds across many of the event types. Examining variations across event types, we find that TSS scores vary much more than IV4 scores. Also, the variation in average TSS scores is consistent with one's priors about these event types. For example, the average scores of analyst recommendations is

[19] Recall that the sum of observations under all event types exceeds the number of observations under "identified days" since they are, on average, multiple events for each identified news day.
[20] Recall that the score ranges from -1 to 1.

close to neutral (0.06) consistent with the idea that analysts revisions are equally likely to be positive as they are to be negative. Legal event, on the other hand are on average negative and correspond to negative TSS scores (-0.21), while partnership events are on average positive and correspond to positive TSS scores (0.63).

These differences between TSS and IV4 scores are not merely an artifact of rescaling. The last column in Table 5 reports the correlation between TSS and IV4 scores. While the correlations are positive and range between 0.18 and 0.38, they are far from one. In fact, for three of the eight event types, event-specific scores correlations are lower than 0.20.

Recall that Table 4 showed a substantial improvement in R^2s once sentiment scores were incorporated into the market model regression framework. In that setting, because of the variable number of observations firm by firm for particular events, the analysis was aggregated to the industry level. Here, we apply the analysis at the firm-by-firm level by constraining the coefficients to be the same across firms. We provide an analysis of the difference between TSS and IV-4 scores as it pertains to our specific interest, stock price relevance. We use weighted least squares regressions (with time clustered errors) and use daily returns as the dependent variable.

These results are reported in Table 6. In the first two columns, we consider TSS and IV4 scores separately. Scores from both methods are positively related to contemporaneous returns, while TSS coefficient is more than twice the size of the IV4 coefficient. Indeed, when we include both scores in a horse race like regression (column 3), the IV4 score is subsumed by TSS score; the IV4 score coefficient becomes statistically insignificant from zero while the size of the TSS coefficient remains large (0.348 vs. 0.355).

In the last column of the Table 6, we combine event identification with sentiment scores. Specifically, we replace daily scores with TSS event-level scores as the independent variables. We find that the contemporaneous relation between event scores and returns, in a multivariate setting, is positive for all event types and significantly different from zero for most of them. The size of the coefficient varies considerably across event types. For example, the analyst recommendation's score coefficient is 1.48 compared with 0.67 for financial events, while both event types have roughly the same distribution of scores. Notably, although the regression R^2s are low, they increase ten-fold when we include event-specific scores. These results confirm the importance of considering differential stock price response to different event types.

V. News Type, Reversals and Continuations

Though the results of Sections 3 and 4 are supportive of one of the main hypotheses from efficient markets, namely, that prices respond to fundamental information, the growing literature in the area of behavioral finance also has implications for our research. There are a number of papers that describe conditions under which stock prices might under- or overreact based on well-documented behavioral biases. (See, for example, Daniel, Hirshleifer and Subrahmanyam (1998), Barberis, Shleifer and Vishny (1998), Hong and Stein (1999), Hirshleifer (2002), and Barber and Odean (2008), among others.) Essential findings from this literature based on behavioral theory are that (i) investors only partially adjust to real information, leading to a continuation of the price response to this information, and (ii) investors overreact to shocks to prices (i.e., unreal information), leading to higher trading volume and reversals of these shocks.

Indeed, there are a number of studies that provide some empirical support for these hypotheses. For example, Stickel and Verrecchia (1994) and Pritamani and Singal (2001) report stock price momentum after earnings announcements. Tetlock, Saar-Tsechansky and Macskassy (2008) report similar underreaction to news events focused on negative words (as measured through a word count based on a textual analysis). The closest papers to ours, however, are Chan (2003) and Tetlock (2010, 2011) who focus on days with and without news. Specifically, Chan (2003) separates out companies hit by price shocks into those with public news versus no news. Chan (2003) finds that after bad news stock prices continue to drift down while after no news stock prices reverse. Tetlock (2010, 2011) generally finds that public news, and especially new as opposed to stale news, reduce the well-known short-term reversals of stock returns. In contrast, Gutierrez and Kelley (2008) do not find any difference.

In this section, we extend the above analyses to our dataset, in particular, to our differentiation of public news into identified news events versus unidentified news. To the extent the behavioral literature tries to explain the theories of under- and overreaction in terms of stock price responses to *real* news versus *false* news, our methodology provides an effective way to study this issue further.

As mentioned above, the results so far suggest a strong contemporaneous response of stocks to their media coverage on identified news days but not on unidentified news days. One interpretation is that identified news days are days on which price-relevant information arrives. To examine this, we measure return autocorrelation on different day types (i.e., no-news, unidentified news, and identified news). Table 7 reports the results of a weighted least squared regression in which the dependent variable is day *t+1* returns. In the first column of the table, the independent variables are time *t* returns and day classification dummies (no-news days dummy is dropped). The results suggest a substantial reversal following no-news days. For example, a 5% return day is followed by, on average, 20bp reversal the next day. This reversal is sizable considering the universe of stocks in our sample and their average bid-ask spreads. Unidentified news days are characterized by reversals too, while the magnitude of the reversals is smaller compared with no-news days (day t return coefficient of -0.022). In contrast, identified news days are followed by continuations (day t return coefficient of 0.010).

Columns 2-9 of Table 7 study these pattern for each of the event types separately. In these regressions, we set the event dummy to be equal to one if the event occurred on date t and zero otherwise. This specification contrasts days on which a specific event took place with all other days. The results suggest that virtually all event types exhibit continuations, with the largest ones following analyst recommendations, employment, financials, and partnerships. Together, the 15 results in the table suggest that the contemporaneous price response to identified news days is unlikely to be due to irrational over-reaction to news coverage of the events underling our study. If anything, it suggests that the price response is insufficiently strong for many of the event types.

Since serial correlation of returns may be non-linear, we compute day *t+1* returns conditional on day *t* return percentile (following the same methodology used in Table 2). Table 8 reports average returns based on the day *t* classification (first three rows) and the day *t* event (next eight rows). The last column in the table reports the difference between the bottom and top 10% columns. The table allows us not only to observe which day types are followed by reversals and which are followed continuations, but it also allows us to detect whether serial correlation in return is coming from negative day returns, positive day returns, or both.

Consistent with the regression results, we find that no news days are characterized by return reversals. The magnitude of the reversals following extreme return days is sizable – 30bp, daily. Unidentified news days are followed by reversals too, while the magnitude is smaller at 18bp. In both cases, most of the reversal comes from negative day returns. In contract, identified day returns are followed by continuations of 20bp, which are driven by positive day returns. We find continuations for seven of the eight event types, with particularly pronounced continuations following analyst recommendations (75bp), employment (41bp), and legal events (65bp).

To further evaluate the economic magnitude of trading on the type of news, we consider two separate zero-cost strategies: (i) follow a reversal strategy following unidentified news days, and (ii) follow a continuation strategy following identified news days. The first strategy goes long (short) stocks with previous day extreme negative (positive) returns and unidentified news, while the second strategy goes long (short) stocks with previous day extreme positive (negative) returns and identified news. Extreme returns are defined as returns that are larger (smaller) than 1.2 (-1.2) times past 20 days' volatility.[21] In all cases we hold the stocks for one day. Table 9 reports the results of these two strategies separately and then of the combined strategy (which puts equal weights on the reversals and

[21] The results are robust to changes in the threshold.

continuation strategies). The top panel of the table reports time series regressions with a four-factor model. The reversals strategy produces an alpha that is indistinguishable from zero.[22] At the same time, the continuations strategy generates a robust alpha of 41bp per day and does not seem to load on any of the factors. Moreover, an analysis of the average returns per year suggests that the strategy produced positive average returns for every year in our sample period, with an overall mean daily return of 46bp.

VI. Conclusions

The bottom line from this paper is in stark contrast to the last 25 years of literature on stock prices and news. We find that, when information can be identified and that the tone (i.e., positive versus negative) of this information can be determined, there is a much closer link between stock prices and information. Examples of results include market model R^2s that are no longer the same on news versus no news days (i.e., Roll's (1988) infamous result), but now are 15% versus 34%; variance ratios of returns on identified news days almost double than those on no news and unidentified news days; and, conditional on extreme moves, stock price reversals occur on no news unidentified news days, while identified news days show continuation.

The methodology described in this paper may be useful for a deeper analysis of the relation between stock prices and information, especially on the behavioral side (e.g., as pertaining to the reversals/continuation analysis of Section 5). There is a vast literature in the behavioral finance area arguing that economic agents, one by one, and even in the aggregate, cannot digest the full economic impact of news quickly. Given our database of identified events, it is possible to measure and investigate "complexity", and its effect on the speed of information processing by the market. For example, "complexity" can be broken down into whether more than one economic event occurs at a given point in time, how news (even similar news) gets accumulated through time, and cross-firm effects of news. We hope to explore some of these ideas in future research.

[22] Consistent with existing research, however, we find that conditioning on no news days, and implementing a reversal strategy, leads to a positive alpha.

References

Aharony, J., and I. Swary, 1980, Quarterly Dividend And Earnings Announcements And Stockholders' Returns: An Empirical Analysis, *Journal of Finance* 35, 1–12.

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, 1259–1293.

Asquith, P., and D. W. Mullins, 1986, Equity Issues And Offering Dilution, *Journal of Financial Economics* 15, 61–89.

Ball, R., and P. Brown, 1968, An Empirical Evaluation Of Accounting Income Numbers, *Journal of Accounting Research* 6, 159–178.

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

Campbell, J. Y., 1991, A Variance Decomposition For Stock Returns, *Economic Journal* 101, 157–179.

Chan, W. S., 2003, Stock Price Reaction To News And No-News: Drift And Reversal After Headlines, *Journal of Financial Economics* 70, 223–260.

Cutler, D. M., J. M. Poterba, and L. H. Summers, 1989, What Moves Stock Prices?, *Journal of Portfolio Management* 15, 4–12.

Das, S. R., and M. Y. Chen, 2007, Yahoo! For Amazon: Sentiment Extraction From Small Talk On The Web, *Management Science* 53, 1375–1388.

Davis, A. K., J. Piger, and L. M. Sedor, 2012, Beyond The Numbers: An Analysis Of Optimistic And Pessimistic Language In Earnings Press Releases, *Contemporary Accounting Research* 29, 845–868.

Demers, E., and C. Vega, 2010, Soft Information In Earnings Announcements: News Or Noise?, *Working Paper, INSEAD.*

Devitt, A., and K. Ahmad, 2007, Sentiment Polarity Identification In Financial News: A Cohesion-

Based Approach, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 984–991.

Engelberg, J., 2008, Costly Information Processing: Evidence From Earnings Announcements, *Working Paper, University of North Carolina*.

Engle, R. F, M. Hansen, and A. Lunde, 2011, And Now, The Rest Of The News: Volatility And Firm Specific News Arrival, *Working Paper*.

Fama, E. F., L. Fisher, M. C. Jensen, and R. Roll, 1969, The Adjustment Of Stock Prices To New Information, *International Economic Review* 10, 1–21.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, ManagementS Tone Change, Post Earnings Announcement Drift And Accruals, *Review of Accounting Studies* 15, 915–953.

Feldman, R., B. Rosenfeld, R. Bar-Haim, and M. Fresko, 2011, The Stock Sonar Sentiment Analysis Of Stocks Based On A Hybrid Approach, *Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference*, 1642–1647.

Feldman, R., and J. Sanger, 2006, The Text Mining Handbook, *Cambridge University Press*.

Griffin, J. M., N. H. Hirschey, and P. J. Kelly, 2011, How Important Is The Financial Media In Global Markets?, *Review of Financial Studies* 24, 3941–3992.

Grob-Klubmann, A., and N. Hautsch, 2011, When Machines Read The News: Using Automated Text Analytics To Quantify High Frequency News-Implied Market Reactions, *Journal of Empirical Finance* 18, 321–340.

Gutierrez, R. C., and E. K. Kelley, 2008, The Long-Lasting Momentum In Weekly Returns, *Journal of Finance* 63, 415–447.

Hanley, K. W., and G. Hoberg, 2012, Litigation Risk, Strategic Disclosure And The Underpricing Of Initial Public Offerings, *Journal of Financial Economics* 103, 235–254.

Kothari, S. P., X. Li, and J. E. Short, 2009, The Effect Of Disclosures By Management, Analysts, And Financial Press On The Equity Cost Of Capital: A Study Using Content Analysis, *Accounting Review* 84, 1639–1670.

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, 2000, Mining Of Concurrent Text And Time Series, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 37–44.

Li, F., 2010, The Information Content Of Forward-Looking Statements In Corporate Filings A Nave Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

Loughran, T., and B. McDonald, 2011, When Is A Liability Not A Liability? Textual Analysis, Dictionaries, And 10-Ks, *Journal of Finance* 66, 35–65.

Mandelker, G., 1974, Risk And Return: The Case Of Merging Firms, *Journal of Financial Economics* 1, 303–335.

Manning, C. D., and H. Schutze, 1999, Foundations Of Statistical Natural Language Processing, *Cambridge Massachusetts: MIT Press.*

Mitchell, M. L., and J. H. Mulherin, 1994, The Impact Of Public Information On The Stock Market, *Journal of Finance* 49, 923–950.

Roll, R., 1984, Orange Juice and Weather, *American Economic Review*, 74, 5, 861-880.

Roll, R., 1988, R2, *Journal of Finance* 43, 541–566.

Shiller, R. J., 1981, Do Stock Prices Move Too Much To Be Justified By Subsequent Changes In Dividends?, *American Economic Review* 71, 421–436.

Tetlock, P. C., 2007, Giving Content To Investor Sentiment: The Role Of Media In The Stock Market, *Journal of Finance* 62, 1139–1168.

Tetlock, P. C., 2010, Does Public Financial News Resolve Asymmetric Information?, *Review of Financial Studies* 23, 3520–3557.

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.

Vega, C., 2006, Stock Price Reaction To Public And Private Information, *Journal of Financial Economics* 82, 103–133.

4 Tables

Table 1: Summary Statistics
Panel A

	# of Obs.	# of Tickers	# of Articles	# of Words	# of Relevant Words
No News	758,393	795	NA	NA	NA
Unid News	334,990	792	2.5	814	107
Iden News	141,720	790	5.9	1,892	412
Total	1,235,103	792	3.5	1,134	198

Panel B

	Stock Return	Market Ret	Size	BM	MOM
No News	0.03%	-0.01%	4.50	2.92	2.85
Unid News	0.05%	0.01%	4.72	2.91	2.83
Iden News	0.05%	0.01%	4.76	2.88	2.82
Total	0.04%	0.00%	4.59	2.91	2.84

The table reports summary statistics for observations (stock/day) classified as having no news, unidentified news (i.e., containing news all with identified events), or identified news (i.e., containing news with some identified events). Panel A reports the total number of observations, the number of unique tickers, the average number of words, and the average number of relevant words (as identified by TSS). Panel B reports the average daily stock return, average daily market return, and the average size, book-to-market, and momentum quintile assignments.

Table 2: Event Frequency Across Return Ranks

Return rank	0-10%	10-30%	30-70%	70-90%	90-100%
# of articles	4.3	3.3	3.2	3.3	4.3
# of words	1,403	1,080	1,054	1,070	1,379
% of rel. words	17.7%	17.4%	17.4%	17.6%	17.3%
No News	-3.5%	1.4%	1.0%	0.9%	-5.2%
Unid News	-5.2%	-0.7%	2.6%	-0.1%	-3.6%
Iden News	26.7%	-5.5%	-9.8%	-3.9%	31.4%
Acquisition	13.6%	-5.2%	-2.8%	-5.5%	18.9%
Analyst Rec	75.8%	-10.8%	-24.8%	-8.5%	62.0%
Deals	3.7%	-6.4%	-0.9%	1.0%	10.7%
Employment	12.7%	-0.5%	-3.9%	-2.2%	8.4%
Financial	59.3%	-9.6%	-21.6%	-9.1%	64.6%
Legal	9.9%	-1.5%	-1.0%	-4.3%	5.7%
Partnerships	-0.2%	-2.1%	0.1%	0.8%	2.2%
Product	3.9%	-3.7%	-1.3%	-1.5%	11.8%

The table reports summary statistics of all observations based on return rank sorts. For each stock separately, we assign each day based on its percentile return rank – bottom 10%, following 20%, middle 40%, following 20%, and top 10%. The statics reported are the average number of article per observation, the average number of words, the fraction of all words identified as relevant (as identified by TSS). Next, we report the difference between the observed distribution and the distribution that would obtain under independence based on observations' classification as having no news, unidentified news (i.e., containing news all with unidentified events), or identified news (i.e., containing news with some identified events). For example, out of a total of 758K no news observations, 75.8K should fall under the bottom 10% of returns, but only 73K do resulting in a -3.5% difference. The bottom panel of the table sorts observations into non-mutually exclusive event types and reports the results of the same comparison described above.

Table 3: Return Variation by Day Type and Events

	Mean	Median	s.d.
No News	7.09	1.25	46.5
Unid News	9.13	1.28	58.3
Iden News	16.14	1.70	124.8
Acquisition	22.08	1.40	238.2
Analyst Rec	35.09	3.01	240.4
Deals	12.91	1.37	122.2
Employment	17.38	1.43	183.1
Financial	22.07	2.42	128.1
Legal	21.23	1.32	256.0
Partnerships	10.82	1.40	150.9
Product	11.27	1.47	111.3
Old news	12.65	1.44	105.6
New news	17.61	1.82	132.0

The table reports daily return variations (daily returns squared) statistics by observation types and event types. Observations (stock/day) are classified as having no news, unidentified news (i.e., containing news all with identified events), or identified news (i.e., containing news with some identified events). Identified news days are flagged by event types (event types are not mutually exclusive per observation).

Table 4: Event R^2s – Firm and Industry-level Regressions with Event Types

Panel A

	Firm Level			Industry Level		
(CAPM)	Mean R^2	Med R^2	N	Mean R^2	Med R^2	N
All	28%	28%	795	28%	28%	60
No News	33%	34%	753	33%	33%	60
Unid News	30%	29%	753	30%	29%	59
Iden News	19%	15%	665	16%	15%	59
TSS Unid News	30%	30%	753	30%	29%	59
TSS Iden News	21%	18%	665	22%	20%	59
IV4 Unid News	30%	30%	753	30%	29%	59
IV4 Iden News	19%	16%	665	20%	18%	59

Panel B

	CAPM		TSS		
	Mean R^2	Med R^2	Mean R^2	Med R^2	N
New News	15%	14%	22%	19%	58
Acquisition	14%	11%	30%	27%	50
Analyst Rec	18%	14%	37%	32%	45
Deals	20%	20%	41%	37%	52
Employment	23%	19%	38%	33%	52
Financial	13%	11%	20%	17%	58
Legal	19%	14%	37%	33%	38
Partnerships	29%	27%	48%	46%	33
Product	29%	28%	43%	41%	40

Panel A of the table reports daily return regressions with one factor (total market, value weighted) in the first four rows, and with two factors in the next four rows (total market and TSS and IV4 sentiment scores). Regressions are run separately for each day category – on all days, no news days, unidentified news days (i.e., containing news all with unidentified events), and identified news days (i.e., containing news with some identified events). Firm level regressions estimate firm-level betas and R^2s while industry level regressions estimate 2-digit SIC industry level betas and R^2s. When sentiment scores are used, daily scores are used for firm level regressions and individual event scores are used for industry level regressions. Panel B of the table reports daily return CAPMP regressions in the first two columns and with two score specific events in the next two columns (total market and TSS sentiment scores). Regressions are run separately for each classification – new news and event type by event type. All industry regressions are pooled at the industry level (2-digit SIC) with WLS and standard errors clustered by time. New news days are defined as days for which at least one of the event types had not appeared in the previous 5 trading days (for the same stock).

Table 5: TSS and IV4 Scores – Summary Statistics

		TSS						IV4						IV4-TSS
	Event count	N	mean	p10	p50	p90	p90-p10	N	mean	p10	p50	p90	p90-p10	Corr.
Unid News	334,990	121,466	0.26	-0.50	0.50	0.75	1.25	327,018	0.34	0.01	0.33	0.70	0.69	0.34
Iden News	141,720	141,641	0.39	-0.39	0.50	0.83	1.22	141,691	0.32	0.08	0.31	0.59	0.51	0.34
Acquisition	19,720	19,684	0.55	0.25	0.50	0.80	0.55	19,689	0.36	0.06	0.37	0.67	0.61	0.19
Analyst Rec	10,584	10,550	0.06	-0.67	0.00	0.75	1.42	10,575	0.23	-0.07	0.22	0.54	0.61	0.33
Deals	27,233	27,215	0.61	0.50	0.67	0.83	0.33	27,218	0.42	0.13	0.43	0.69	0.56	0.19
Employment	18,988	18,982	0.43	-0.40	0.50	0.80	1.20	18,974	0.30	-0.06	0.34	0.64	0.70	0.18
Financial	61,348	61,276	0.29	-0.53	0.50	0.85	1.38	61,304	0.29	-0.01	0.29	0.62	0.63	0.38
Legal	9,126	9,123	-0.21	-0.80	-0.40	0.57	1.37	9,125	0.17	-0.16	0.17	0.50	0.66	0.31
Partnerships	9,301	9,295	0.63	0.50	0.67	0.83	0.33	9,295	0.44	0.16	0.46	0.69	0.53	0.24
Product	23,175	23,162	0.56	0.25	0.67	0.83	0.58	23,153	0.35	0.00	0.38	0.65	0.65	0.31

The table reports summary statistics of scores computed based on TSS and IV4 identification of positive (P) and negative (N) features. Scores in all cases equal to $\frac{P-N}{P+N+1}$. Observations are sorted by day type: unidentified news (i.e., containing news all with unidentified events), or identified news (i.e., containing news with some identified events), and event types. The first column reports the total number of observations under each day type. The column block titled "TSS"("IV4") reports results obtained from TSS(IV4). N corresponds to the number of observations with non-missing scores, and the remaining columns report the distribution of scores. The last column in the table reports the correlation between TSS and IV4 scores for each of the days types.

Table 6: TSS and IV4 Scores – Regression Analysis

	IV4	TSS	IV4-TSS	TSS
Daily score IV4	0.145		0.045	
	[0.029]***		[0.028]	
Daily score TSS		0.355	0.348	
		[0.017]***	[0.016]***	
Acquisition score				0.059
				[0.096]
AnalystRec score				1.483
				[0.082]***
Deals score				0.125
				[0.095]
Employment score				0.147
				[0.052]***
Financial score				0.672
				[0.039]***
Legal score				0.146
				[0.056]***
Partner score				0.193
				[0.132]
Product score				0.169
				[0.064]***
$I_{IV4score}$	-0.031		-0.003	
	[0.017]*		[0.016]	
$I_{TSSscore}$		-0.103	-0.109	
		[0.015]***	[0.012]***	
$I_{Acquisitionscore}$				0.017
				[0.062]
$I_{AnalystRecscore}$				-0.249
				[0.051]***
$I_{Dealsscore}$				-0.072
				[0.065]
$I_{Employmentscore}$				-0.143
				[0.040]***
$I_{Financialscore}$				-0.252
				[0.033]***
$I_{Legalscore}$				-0.031
				[0.035]
$I_{Partnerscore}$				-0.179
				[0.087]**
$I_{Productscore}$				-0.124
				[0.043]***
Constant	0.039	0.042	0.039	0.106
	[0.024]	[0.024]*	[0.024]	[0.033]***
Observations	1,235,103	1,235,103	1,235,103	1,235,103
R^2	0.000	0.001	0.001	0.011

The dependent variable in all specifications are day t stock returns. The independent variables include daily scores and event specific scores based on TSS and IV4 identification of positive (P) and negative (N) features. Scores in all cases equal to $\frac{P-N}{P+N+1}$. All regressions use WLS with time clustered standard errors.

Table 7: Return Reversals and Continuations

(Ret_1)	By day type	Acquisition	AnalystRec	Deals	Employment	Financial	Legal	Partner	Product
Ret_0	-0.038	-0.024	-0.026	-0.026	-0.026	-0.03	-0.025	-0.025	-0.024
	[0.014]***	[0.012]**	[0.012]**	[0.012]**	[0.012]**	[0.013]**	[0.012]**	[0.012]**	[0.012]**
I_{unid} x Ret_0	0.016								
	[0.007]**								
I_{iden} x Ret_0	0.048								
	[0.009]***								
I_{event} x Ret_0		-0.004	0.046	0.042	0.058	0.045	0.038	0.042	0.003
		[0.018]	[0.022]**	[0.015]***	[0.018]***	[0.010]***	[0.038]	[0.020]**	[0.011]
I_{unid}	0.007								
	[0.013]								
I_{iden}	-0.005								
	[0.016]								
I_{event}		0.018	-0.039	0.001	-0.024	-0.005	-0.015	-0.02	-0.026
		[0.021]	[0.036]	[0.017]	[0.021]	[0.020]	[0.027]	[0.028]	[0.021]
Constant	0.046	0.047	0.048	0.048	0.048	0.048	0.048	0.048	0.048
	[0.025]*	[0.024]**	[0.024]**	[0.024]**	[0.024]**	[0.024]**	[0.024]**	[0.024]**	[0.024]**
Observations	1,234,308	1,234,308	1,234,308	1,234,308	1,234,308	1,234,308	1,234,308	1,234,308	1,234,308
R^2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

The dependent variable in all regressions is day $t+1$ stock returns. The dependent variables include day t stock returns, in all specifications, and day t classification dummy: I_{unid} equals one for days with unidentified news, I_{iden} equals one for days with identified news, and I_{event} equals one for days with the column event type. All regressions use WLS with time clustered standard errors.

Table 8: Reversals and Continuations – Past Return Sorts

Return rank	0-10%	10-30%	30-70%	70-90%	90-100%	Differ
No News	0.25	0.04	0.03	0.00	-0.05	0.30
Unid News	0.18	0.04	0.04	0.02	0.00	0.18
Iden News	-0.07	0.01	0.03	0.02	0.13	-0.20
Acquisition	0.08	0.08	0.01	-0.01	0.20	-0.12
Analyst Rec	-0.27	-0.01	0.00	-0.10	0.48	-0.75
Deals	-0.10	0.03	0.05	0.08	0.08	-0.18
Employment	-0.18	-0.03	0.01	-0.01	0.23	-0.41
Financial	-0.13	-0.03	0.04	0.05	0.14	-0.27
Legal	-0.25	0.00	0.04	-0.07	0.39	-0.65
Partnerships	0.04	-0.02	0.01	-0.05	0.28	-0.24
Product	0.02	0.04	0.04	0.02	-0.14	0.16

The table reports day $t+1$ stock returns. Observations are sorted based on day t return ranks and day t classification. Return ranks are computed as follows: for each stock separately, we assign each day based on its percentile return rank – bottom 10%, following 20%, middle 40%, following 20%, and top 10%. Observations are also classified as having no news, unidentified news (i.e., containing news all with identified events), or identified news (i.e., containing news with some identified events), and by event types. All figures are in percentage terms.

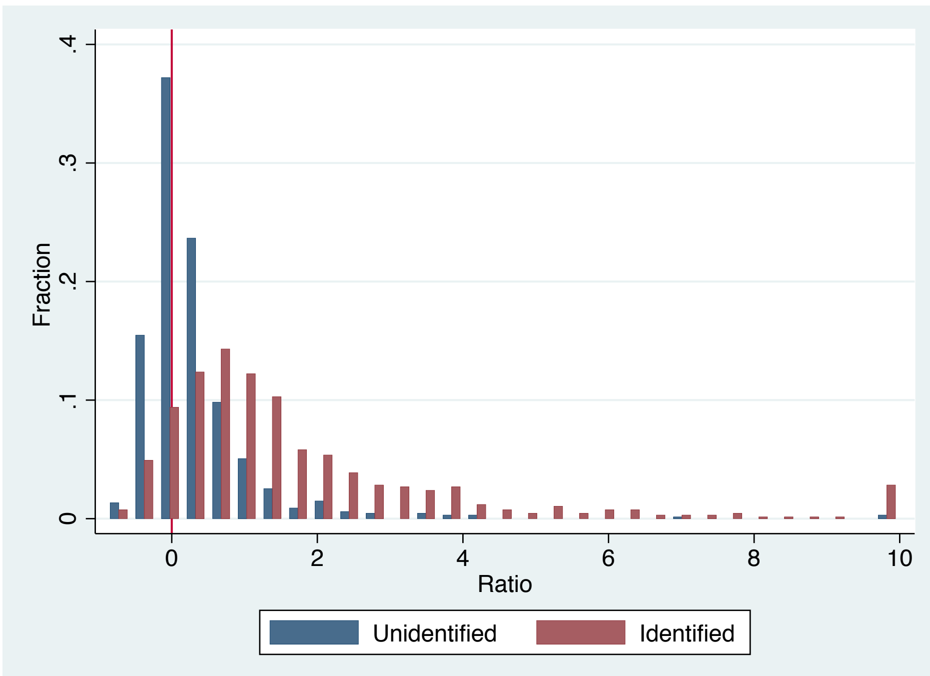
Table 9: Reversals and Continuations Strategies

	Reversals	Continuations	Combined
Alpha	-0.023	0.405	0.155
	[0.065]	[0.109]***	[0.074]**
Mkt-rf	0.265	-0.322	0.015
	[0.069]***	[0.201]	[0.103]
SMB	-0.121	0.376	0.107
	[0.134]	[0.314]	[0.171]
HTM	-0.093	-0.132	-0.022
	[0.119]	[0.258]	[0.141]
UMD	0.136	-0.024	0.102
	[0.080]*	[0.119]	[0.077]
Observations	2181	2111	2321
R^2	0.013	0.01	0.002

Year		Mean daily return	
2000	0.131%	0.228%	0.219%
2001	-0.216%	1.381%	0.539%
2002	0.082%	0.982%	0.592%
2003	-0.118%	0.073%	-0.128%
2004	0.086%	0.307%	0.186%
2005	0.011%	0.231%	0.116%
2006	-0.067%	0.116%	0.009%
2007	-0.032%	0.295%	0.131%
2008	-0.454%	1.017%	0.139%
2009	0.280%	0.219%	0.308%
Total	-0.030%	0.463%	0.198%

The table reports zero-cost trading strategy returns based on day $t-1$ classified: unidentified news (i.e., containing news all with unidentified events) and identified news (i.e., containing news with some identified events). Reversals strategy (first column) goes long(short) stocks classified on day $t-1$ as having large negative(positive) returns and no news. Likewise, continuation strategy (second column) goes long(short) stocks classified on day $t-1$ as having large positive(negative) returns and identified news. Combined strategy invests equally in reversals and continuations. Large positive(negative) returns are defined those exceeding(falling below) own stocks' 1.2×20 day lagged volatility. Holding period for all strategies is one day. The top panel reports four factor time series regressions. The bottom panel reports average daily returns from the strategies for each of the years in our sample period.

Figure 1: Variance Ratios



The figure reports the distribution of the ratios between squared daily returns on unidentified news days and no news days, and the ratios between daily squared return on identified news days and no news days (minus 1). Unidentified news are days containing news all with identified events and identified news are days containing news with some identified events. Ratios are winsorized at 10.