OFR WORKING PAPER
OFFICE OF FINANCIAL RESEARCH

16-04 | April 20, 2016

Does Unusual News Forecast Market Stress?

Harry Mamaysky
Columbia University
hm2646@columbia.edu

Paul Glasserman
Office of Financial Research and Columbia University
paul.glasserman@ofr.treasury.gov

Does Unusual News Forecast Market Stress?

Harry Mamaysky and Paul Glasserman*

Abstract

We find that an increase in the "unusualness" of news with negative sentiment predicts an increase in stock market volatility. Our analysis is based on more than 360,000 articles on 50 large financial companies, mostly banks and insurers, published in 1996–2014. We find that the interaction between measures of unusualness and sentiment forecasts volatility at both the company-specific and aggregate level. These effects persist for several months. The pattern of response of volatility in our aggregate analysis is consistent with a model of rational inattention among investors.

*Columbia Business School and Office of Financial Research. This paper was produced while Paul Glasserman was a consultant to the OFR. We acknowledge the excellent research assistance of Il Doo Lee. We thank Paul Tetlock, Geert Bekaert, and participants at the Summer 2015 Consortium for Systemic Risk Analytics conference, Columbia, the Office of Financial Research, the High Frequency Finance and Analytics conference at the Stevens Institute, and the IAQF/Thalesians seminar for valuable comments. We thank the Thomson Reuters Corp. for graciously providing the data that was used in this study. We use the Natural Language Toolkit in Python for all text processing applications in the paper. For empirical analysis we use the **R** programming language for statistical computing.

1 Introduction

Can the content of news articles forecast market stress and, if so, what type of content is predictive? Several studies have documented that news sentiment forecasts market returns. We find that a measure of "unusualness" of news text combined with sentiment forecasts stress, which we proxy by stock market volatility. The effects we find play out over months, whereas in most prior work the stock market's response to news articles dissipates in a few days.

The link between sentiment expressed in public documents and stock market returns has received a great deal of attention. At an aggregate level, Tetlock (2007) finds that negative sentiment in the news depresses returns; Tetlock, Saar-Tsechansky, and Macskassy (2008) study company-specific news stories and responses. Garcia (2008) finds that the influence of news sentiment is concentrated in recessions. Loughran and McDonald (2011) and Jegadeesh and Wu (2013) apply sentiment analysis to 10-K filings. Da, Engelberg, and Gao (2014) measure sentiment in Internet search terms. Manela and Moreira (2015) find that a news-based measure of uncertainty forecasts returns. Our focus differs from prior work because we seek to forecast market stress rather than the direction of the market. We apply new tools to this analysis, going beyond sentiment word counts.

The importance of unusualness is illustrated by the following two phrases, both of which appeared in news articles from September 2008:

> "the collapse of Lehman"
> "cut its price target"

Both phrases contain one negative word and would therefore contribute equally to an overall measure of negative sentiment in a standard word-counting analysis. But we recognize the first phrase as much more unusual than the second, relative to earlier news stories. This difference can be quantified by taking into account the frequency of occurrence of the phrases in prior months. As this simple example suggests, we find that sentiment is important, but it becomes more informative when interacted with our measure of unusualness.

Research in finance and economics has commonly measured sentiment through what is known in the natural language processing literature as a bag-of-words approach: an article is classified as having positive or negative sentiment based on the frequency of positive or negative connotation words that it contains. The papers cited above are examples of this approach. As the example above indicates, this approach misses important information: the unusualness of the first phrase lies not in its use of “collapse” or “Lehman” but in their juxtaposition. We therefore measure unusualness of consecutive word phrases rather than individual words.

Our analysis uses all news articles in the Thomson Reuters Corp. database between January 1996 and December 2014 that mention any of the top 50 global banks, insurance, and real estate firms by market capitalization as of February 2015. After some cleaning of the data, this leaves us with 367,331 articles, for an average of 1,611 per month. We calculate measures of sentiment and unusualness from these news stories and study their ability to forecast realized or implied volatility at the company-specific and aggregate levels.

The consistent picture that emerges from this analysis is that the interaction of unusualness with negative sentiment yields the best predictor of future stock market volatility among the news measures we study. Also, our analysis shows that news is not absorbed by the market instantaneously.

We first run forecasting regressions of company-specific implied volatility on company-specific news measures and evaluate results in the cross section. Our interacted measure of unusual negative news, $ENTSENT\_NEG$, provides a statistically and economically significant predictor of volatility at lags of up to six months. Negative sentiment and unusualness are also significant separately, but much less so. Across companies, our interacted measure increases R^2 by an average of 0.22, relative to a baseline control model, more than any of the other news measures we consider.

We then introduce controls for lagged values of implied and realized volatilities and negative returns, all of which are known to be relevant predictors of volatility; see for example Bekaert and Hoerova (2014). We include these controls in panel regressions of company-specific volatility measures on company-specific news measures. Our interacted measures of sentiment (both positive and negative) and unusualness remain economically and statistically significant, even with the inclusion of the controls, with positive measures forecasting a decrease in volatility and negative measures forecasting an increase. These results indicate that the information in our news measures is not fully reflected in contemporaneous prices. In regressions of individual companies, the incremental R^2 from our news measures is much smaller once we control for lagged volatility, but it remains largest for the interacted measure $ENTSENT\_NEG$.

For our aggregate analysis we extract aggregate measures of unusualness and sentiment from our full set of news articles. We estimate vector autoregressions, taking as state variables the VIX, realized volatility on the S&P 500, and several aggregate news measures. We examine interactions among the variables through impulse response functions. A shock to either negative sentiment or our interacted variable $ENTSENT\_NEG$ produces a statistically significant increase in both implied and realized volatility over several months. Once again, the effect is strongest for our interacted measure of unusual negative news. The response of implied and realized volatility to an impulse in $ENTSENT\_NEG$ (or negative sentiment) is hump-shaped, peaking at around four months. This pattern suggests that the information in these news variables is absorbed slowly, a point we return to later.

As a final test of the informativeness of our news measures, we evaluate the performance of long-short portfolios based on sorting stocks on sentiment and unusualness measures calculated from company-specific news articles. We test two sorts on sentiment — one using just negative sentiment and one using both positive and negative sentiment — and two sorts based on interacting these sentiment variables with unusualness. We test several holding periods and fractions of stocks included in the portfolios. With only 50 stocks to work with, our results are sensitive to these parameters, but in most cases sorting on interacted variables produces higher excess returns (relative to market, size, value, and momentum factors) and higher Sharpe ratios than sorting on sentiment alone.

In most prior work that finds a predictive signal in the text of public documents, the information is incorporated into prices within a few days.[1] In contrast, we find that news measures forecast volatility at lags as long as six months. These longer-horizon effects are intuitively plausible in forecasting volatility rather than the direction of the market. Brewing concerns often generate public discourse well before they materialize as market stress (if they do materialize). For example, Google Trends data shows searches for "subprime" spiking in March 2007, more than three months before the sharp rise in market volatility in July 2007. Concerns about a Greek exit from the euro have been in the news for years, yet there is little doubt that the event itself would drive up volatility, despite the anticipation. Arbitraging a predictable rise in volatility is much more difficult than profiting from a predictable stock return: the term structure of implied volatility is typically upward sloping, the roll yield on VIX futures is typically negative, and implied volatility is typically higher than realized volatility, so trades based on options, futures or variance swaps need to overcome these hurdles. The uncertainty around whether a potential stress will materialize (think of betting on Y2K fears) may further dampen risk-adjusted returns from trading on forecasted volatility.

Rational inattention offers a possible explanation for the patterns we observe. Several studies

[1] An exception is Heston and Sinha (2014). By aggregating news weekly, they find evidence of predictability over a three-month horizon.

have found evidence that the limits of human attention affect market prices; see, for example, the survey of Daniel, Hirshleifer, and Teoh (2004). Models of rational inattention, as developed in Sims (2003, 2015), attach a cost or constraint on information processing capacity: investors cannot (or prefer not to) spend all their time analyzing the price implications of all available information. We interpret the cost or constraint on information processing broadly. It includes the fact that people cannot read thousands of news articles per day (and having a computer do the analysis involves some investment); but it also reflects limits on the contracts investors can write to hedge market stress, given imperfect information on unobservable macro state variables. Even among professionals, many investors may focus on a narrow set of stocks or industries and may overlook information that becomes relevant only when aggregated over many stocks. Indeed, Jung and Shiller (2005) review empirical evidence supporting what they call Samuelson's dictum, that the stock market is micro efficient but macro inefficient. The allocation of attention between indiosyncratic and aggregate information drives the model of Maćkowiak and Wiederholt (2009b). Investors also need to allocate attention across different time horizons. Dellavigna and Pollet (2007) find that demographic information with long-term implications is poorly reflected in market prices. A related effect may apply in our setting: investors may anticipate the possibility of elevated volatility in the future yet not take actions that eliminate this outcome.

Beyond this qualitative link to rational inattention we develop a precise connection. First, we argue that although investors would like to hedge aggregate risk, information constraints make it impossible to write contracts directly tied to unobservable macro state variables. We interpret the VIX as an example of a resulting imperfect hedge. Next we evaluate the price of an approximate hedge in a formulation consistent with rational inattention, meaning that investors evaluate the conditional expectation of future cash flows based on imperfect information about the past. Building on work of Sims (2003, 2015) and Maćkowiak and Wiederholt (2009b), we show that when investors face binding information-processing constraints, the response of the VIX to an impulse in the macro state variable is hump-shaped rather than monotonic, consistent with what we find in our vector autoregressions. In other words, information constraints cause news about macro shocks to be incorporated in the VIX only gradually.

Because the effects we find in the data play out over months, the signals we extract from news articles are potentially useful for monitoring purposes. Along these lines, Baker, Bloom, and Davis (2013) develop an index of economic policy uncertainty based (partially) on newspaper articles. Indicators of systemic risk (see Bisias et al. 2012) are generally based on market prices or lagged economic data; incorporating news analysis offers a potential direction for improved monitoring of stress to the financial system. From a methodological perspective, our work applies two ideas from the field of natural language processing to text analysis in finance. As already noted, we measure the "unusualness" of language, and we do this through a measure of entropy in word counts. Also, we take consecutive strings of words (called n-grams) rather than individual words as our basic unit of analysis. In particular, we calculate the unusualness (entropy) of consecutive four-word sequences. These ideas are developed in greater detail in Jurafsky and Martin (2009).

The rest of this paper is organized as follows. Section 2 introduces the methodology we use, and Section 3 discusses the empirical implementation. Section 4 presents results based on company-specific volatility, and Section 5 examines aggregative volatility. Section 6 looks at return predictability using unusualness and sentiment. Section 7 develops the connection with rational inattention. Section 8 concludes.

2 Methodology

2.1 Unusualness of language

A text is unusual if it has low probability, so measuring unusualness requires a model of the probability of language. This problem has been studied in the natural language processing literature on word prediction. Jurafsky and Martin (2009), a very thorough reference for the techniques we employ in this paper, gives the following example: What word is likely to follow the phrase *please turn your homework ...*? Possibly it could be *in* or *over*, but a word like *the* is very unlikely. A reasonable language model should give a value for

$$P(in|please\ turn\ your\ homework)$$

that is relatively high, and a value for

$$P(the|please\ turn\ your\ homework)$$

that is close to zero. One way to estimate these probabilities is to count the number of times that *in* or *the* have followed the phrase *please turn your homework* in a large body of relevant text.

To use an example from our dataset, up until October 2011, which is around the start of the European sovereign debt crisis, the phrase *negative outlook on* had appeared 688 times, and had always been followed by the word *any*. In October 2011, we observe in our sample 13 occurrences of the phrase *negative outlook on France*. We would like our language model to consider this phrase unusual given the observed history.

An *n-gram* is a sequence of n words or, more precisely, n tokens.[2] Models that compute these types of probabilities are called n-gram models (in this example, $n = 5$) because they give the probability of seeing the fifth word conditional on the first four for a given 5-gram.

Consider the N-word text $w_1 \dots w_N$. We can write its probability as

$$P(w_1 \dots w_N) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_N|w_1w_2 \dots w_{N-1}). \tag{1}$$

N-gram models are used in this context to approximate conditional probabilities of the form $P(w_k|w_1 \dots w_{k-1})$ when k is so large (practically speaking, for $k \geq 6$) that it becomes difficult to provide a meaningful estimate of the conditional probabilities for most words. In the case of an n-gram model, we would approximate the above with

$$P(w_k|w_1 \dots w_{k-1}) \approx P(w_k|w_{k-(n-1)} \dots w_{k-1}),$$

which allows us to approximate the probability in (1) as

$$P(w_1 \dots w_N) = \prod_{k=n}^{N} P(w_k|w_{k-n+1} \dots w_{k-1}). \tag{2}$$

In (2), we have dropped the probability $P(w_1 \dots w_{n-1})$ of the first $n-1$ words, which should have little effect if $n \approx 4$ and N is in the thousands.

Let us refer to the text whose probability (or unusualness) we are trying to determine as the *evaluation text*. Since the true text model is not known, the probabilities in (2) will usually have to be estimated from a *training corpus*, $\tilde{w}_1 \dots \tilde{w}_{\tilde{N}}$, where typically $\tilde{N} \gg N$. The idea is to use a large collection of text to estimate the probability that a given word will follow a certain phrase, and then to use these conditional probabilities to determine a probability score for text that we encounter later on.

[2] For example, we treat "chief executive officer" as a single token. When we refer to "words" in the following discussion, we always mean tokens.

Consider an evaluation text $w_1 \dots w_N$ and conditional probabilities $P(w_k | w_{k-n+1} \dots w_{k-1})$ estimated from a training corpus $\tilde{w}_1 \dots \tilde{w}_{\tilde{N}}$.[3] Assuming there are I distinct n-grams in $w_1 \dots w_N$, we can reorganize (2) as

$$P(w_1 \dots w_N) = \prod_{i=1}^{I} P(\omega_n^i | \omega_1^i \cdots \omega_{n-1}^i)^{c_i}, \tag{3}$$

where $\{\omega_1^i \dots \omega_n^i\}$ is the i^{th} n-gram, and c_i is the number of times this n-gram appears in the evaluation text $w_1 \dots w_N$ (so that $c_1 + \cdots + c_I = N - n + 1$).

The probabilities $P(\omega_n^i | \omega_1^i \cdots \omega_{n-1}^i)$ in (3) are estimated from the training corpus. For a 4-gram $\{\omega_1\ \omega_2\ \omega_3\ \omega_4\}$, the empirical probability of ω_4^i conditional on $\omega_1\ \omega_2\ \omega_3$ will be denoted by m_i, and is given by

$$m_i = \frac{c(\{\omega_1\ \omega_2\ \omega_3\ \omega_4\})}{c(\{\omega_1\ \omega_2\ \omega_3\})} \tag{4}$$

where $c(\cdot)$ is the count of the given 3- or 4-gram in the training corpus.

Taking logs in (3) and dividing by the total number of n-grams in the evaluation text, $w_1 \dots w_N$, we obtain the per word, negative log probability of this text:

$$\begin{aligned} H(w_1 \dots w_N) &\equiv -\frac{1}{N-n+1} \log P(w_1 \dots w_N) = -\frac{1}{N-n+1} \sum_{i=1}^{I} c_i \log m_i \\ &= -\sum_{i=1}^{I} p_i \log m_i, \end{aligned} \tag{5}$$

where p_i is the frequency of occurrence of n-gram i in the evaluation text $w_1 \dots w_N$.

The evaluation text $w_1 \dots w_N$ is *unusual* if it has low probability $P(w_1, \dots, w_N)$, relative to the training corpus. Equation (5) shows that, in an n-gram model, the evaluation text is unusual if there are n-grams i that occur frequently in the evaluation text (as measured by p_i) but rarely in the training corpus (as measured by m_i).

The quantity in (5) is called the cross-entropy of the model probabilities m_i with respect to the observed probabilities p_i (see Jurafsky and Martin (2009) equation (4.62)). We refer to $H(w_1 \dots w_N)$ simply as the entropy of the evaluation text. Based on this definition, unusual

[3] We will address in Section 3.3 how to handle the situation that the n-gram $\{w_{k-n+1} \dots w_k\}$ was not observed in the training corpus.

texts will have high entropy.[4]

Lists of n-grams

The definition of entropy in (5) applies to an arbitrary list of n-grams, as opposed to just a text $w_1 \dots w_N$, as long as we know the count c_i for each n-gram i. For example, we may want to consider the list of n-grams that include the word "France," or the list of all n-grams appearing in articles about banks. For a list j of n-grams, we denote by $\{c_1^j(t), \dots, c_I^j(t)\}$ the counts of the number of times each n-gram appears in month t. For an n-gram i that does not appear in list j in month t, $c_i^j(t) = 0$.

Given these counts, for each n-gram i we can calculate

$$p_i^j(t) = \frac{c_i^j(t)}{\sum_i c_i^j(t)},$$

which is i's fraction of the total count of n-grams in list j. Given a list of n-grams in month t, the entropy of that list will be defined as

$$H^j(t) \equiv -\sum_i p_i^j(t) \log m_i(t), \tag{6}$$

which is a generalization of (5). As in (4), the m_i are conditional probabilities estimated from a training corpus. We write $m_i(t)$ to emphasize that these are estimated from news articles published prior to month t. As explained in Section 3.3, we use a rolling window to calculate $m_i(t)$.

2.2 Sentiment

The traditional approach for evaluating sentiment has been to calculate the fraction of words in a given document that have negative or positive connotations.[5] To do so, researchers rely

[4] Tetlock (2011) uses measures of similarity between news articles as proxies for staleness of news, and the same tools could potentially be used to measure dissimilarity as a proxy for unusualness. Tetlock's (2011) approach seems better suited to comparing pairs of articles than to comparing large bodies of text.

[5] Loughran and McDonald (2011) use a more sophisticated approach that assigns higher weights to negative or positive sentiment words that occur less frequently in a training corpus. Jegadeesh and Wu (2013) empirically assess the importance of words by regressing contemporaneous returns of companies releasing 10K's on the frequency of occurrence of words in those filings.

on dictionaries that classify words into different sentiment categories. Tetlock (2007) and Tetlock, Saar-Tsechansky, and Macskassy (2008) use the Harvard IV-4 psychosocial dictionary. Recent evidence (Loughran and McDonald (2011) and Heston and Sinha (2014)) shows that the Laughran-McDonald[6] word lists do a better job of sentiment categorization in a financial context than the Harvard dictionary. We use the Laughran-McDonald dictionary in our work.

Because our core unit of analysis is the n-gram, we take a slightly different approach than the traditional literature. Rather than counting the number of positive or negative words in a given article, we classify n-grams as being either positive or negative. An n-gram is classified as positive (negative) if it contains at least one positive (negative) word and no negative (positive) words. We can then measure the tone of (subsets of) news stories by looking at the fraction of n-grams they contain which are classified as either positive or negative.

3 Empirical implementation

Our dataset consists of Thomson Reuters news articles about the top 50 global banks, insurance, and real estate firms by U.S. dollar market capitalization as of February 2015.[7] Almost 90 percent of the articles are from Reuters itself, with the remainder coming from one of 16 other news services. Table 1 lists the companies in our sample. Table 2 groups our sample of companies and articles by country of domicile. The table reports the following statistics about companies domiciled in a given country: (1) average market capitalization, (2) the percent of all articles that mention companies from that country, and (3) the number of companies. Our set of news articles leans heavily towards the English speaking countries (US, UK, Australia, Canada). For example, even though China has 8 (of a total of 50) companies with market capitalizations on par with the U.S. companies, under 3 percent of our total articles mention companies from China.

The raw dataset has over 600,000 news articles, from January 1996 to December 2014. Many articles represent multiple rewrites of the same initial story. We filter these by keeping only the first article in a given chain.[8] We also drop any article coming from PR Newswire, as these are corporate press releases. All articles whose headlines start with `REG-` (regulatory filings) or `TABLE-` (data tables) are also excluded. This yields 367,331 unique news stories which we

[6]See `http://www3.nd.edu/~mcdonald/Word_Lists.html`.

[7]The survivorship bias in this selection of companies works against the effects we find — firms that disappeared during the financial crisis are not in our sample.

[8]All articles in a chain share the same *Reuters ID* code.

ultimately use in our analysis. Each article is tagged by Thomson Reuters with the names of the companies mentioned in that article. Many articles mention more than one company. Section A.1 gives more details about our data processing.

Figure 1 shows the time series of article counts in our sample. The per month article count reaches its approximate steady-state level of 1,500 or so articles in the early 2000's, peaks around the time of the financial crisis, and settles back down to the steady state level towards the end of 2014. The early years of our sample have relatively fewer articles, which may introduce some noise into our analysis.

Our market data comes from Bloomberg L.P. For each of the 50 companies in our sample we construct a U.S. dollar total returns series using Bloomberg price change and dividend yield data. Also, for those firms that have traded options, we use 30-day implied volatilities for at-the-money options from the Bloomberg volatility surfaces. Our macro data series are the Chicago Board Options Exchange Volatility Index (VIX) and 30-day realized volatility for the S&P 500 Index computed from daily returns.[9]

Throughout the paper, our empirical work is at a monthly horizon, both for our news measures and our market and volatility data.

3.1 N-grams

In our empirical work, we use a 4-gram model.[10]

Each article goes through a data-cleaning process to yield more meaningful n-grams. For example, all company names (and known variations) are replaced with the string *_company_*. Phrases such as *Goldman Sachs reported quarterly results* and *Morgan Stanley reported quarterly results* are replaced with *_company_ reported quarterly results* thus reducing two distinct 4-grams into a single one that captures the semantic intent of the originals. In this way we reduce the number of n-grams in our sample, which will allow us to better estimate conditional probabilities in our training corpus. In another example, we replace *chief executive officer* with *ceo* because we would like the entity referred to as *ceo* to appear in n-grams as a single token, rather than a three word phrase. Appendix A.1 gives more details about our cleaning procedure.

[9]Month t realized returns are returns realized in that month, whereas the month t VIX level is the close-of-month level.

[10]Jurafsky and Martin (2009, p. 112) discuss why 4-gram models are a good choice for most training corpora.

We collect all 4-grams that appear in cleaned articles.[11] An n-gram must appear entirely within a sentence. Contiguous words that cross sentences do not count as an n-gram.[12] For month t we consider various lists of n-grams, such as the list of all n-grams appearing in time t articles, or the list of n-grams that appear in time t articles that mention a specific company.

For example, in January of 2013, the 4-gram *raises target price to* appeared 491 times in the entire sample (i.e. $c^{All}_{\{raises\ target\ price\ to\}}$(January 2013) = 491 where All is the list of n-grams appearing in all articles). It appeared 34 times in articles that were tagged as mentioning Wells Fargo & Co. 26 times in articles that mentioned JPMorgan Chase & Co., but 0 times in articles that mentioned Bank of America Corp. If we sum across all 50 names in our dataset, this 4-gram appeared 1,014 times (more than its total of 491 because many articles mention more than one company).

In each month, we focus on the 5000 most frequently occurring 4-grams. In our 19 year dataset, we thus analyze $19 \times 12 \times 5000 = 1.14$mm 4-grams. Of these 4-grams, 394,778 are distinct. The first three tokens in the latter represent 302,973 distinct 3-grams.

3.2 Sentiment

We define sentiment of a given subset of articles as the percentage of the total count of all n-grams appearing in those articles that are classified as either positive or negative. For example, we may be interested in those articles mentioning Bank of America, or JPMorgan, or the set of all articles at time t. If we denote by $POS(t)$ ($NEG(t)$) the set of all time t n-grams that are classified as positive (negative), then the positive sentiment of list j is

$$SENTPOS^j(t) = \frac{\sum_{i \in POS(t)} c^j_i(t)}{\sum_i c^j_i(t)}, \tag{7}$$

with the analogous definition for $SENTNEG^j(t)$.

For the list of n-grams from all time t articles, we will simply omit the superscript. For all n-grams coming from articles that mention, say, JPMorgan we would write $SENTPOS^{JPM}(t)$. Figure 2 shows the time series of $SENTPOS$ and $SENTNEG$ in our sample, as well as a scaled version of the VIX. Note that at the aggregate level, negative sentiment appears to be

[11] We use the Natural Language Toolkit package in Python for all text processing applications in the paper (see Section A.1).

[12] Note that this imposes slightly more structure than what is assumed about $w_1 \dots w_N$ in (1).

contemporaneously positively correlated with the VIX, whereas positive sentiment is contemporaneously negatively correlated. The correlations are 0.458 and -0.373 respectively. Section 5 will study the dynamics of this relationship in depth.

Table 3 shows the average contemporaneous correlation between the 50 individual implied volatility and sentiment pairs (i.e. between single name implied volatility and the $SENTNEG^j$ and $SENTNEG^j$ series for a given company j), and between the aggregate sentiment series and the VIX. If an individual implied volatility series does not exist, we use the VIX instead. Cross-sectional standard errors are also calculated assuming independence of observations. Averaging across single names reveals that $SENTNEG^j$ ($SENTPOS^j$) is on average positively (negatively) correlated with single name implied volatility, which is consistent with what we observe at the aggregate level.

We thus have fairly strong evidence that our sentiment measures, at the aggregate and single name levels, are responding to the same factors that drive the VIX.

3.3 Entropy

Our entropy measures come from equation (6). We refer to the measure of unusualness of all time t articles as $ENTALL(t)$. The unsualness of only those articles which mention a specific company is $ENTALL^j(t)$, where j is the list of n-grams coming from articles that mention the company in question.

We can also measure the unusualness of subsets of n-grams that do not correspond to all n-grams that come from some set of articles. For example, we can look at the list of n-grams which are classified as having negative (positive) sentiment; we refer to this entropy measure as $ENTNEG(t)$ ($ENTPOS(t)$). Or we can look at the list of n-grams that have negative (positive) sentiment that come from the subset of articles in month t that mention company j; we refer to these measures as $ENTNEG^j(t)$ ($ENTPOS^j(t)$).

N-grams from month t articles form the evaluation text (giving us the p_i^j's), and n-grams from rolling windows over past articles form the training corpus (giving us the m_i's). The training corpus for month t consists of all 3- and 4-grams in our dataset that appeared in the two year period from month $t-27$ up to and including month $t-4$. We use a rolling window, as opposed to an expanding window from the start of the sample to $t-4$ in order to keep the information sets for all our entropy calculations of roughly the same size.

It is possible that a given 4-gram that we observe in month t never occurred in our sample prior to month t. In this case $m_i(t)$ is either zero (so its log is infinite) or undefined if its associated 3-gram also has never appeared in the training sample. To address this problem, we modify our definition of $m_i(t)$[13] in (4) to be

$$m_i(t) \equiv \frac{c(\{\omega_1\omega_2\omega_3\omega_4\}) + 1}{c(\{\omega_1\omega_2\omega_3\}) + 4}. \tag{8}$$

This means that a 4-gram/3-gram pair that has never appeared in our sample prior to t will be given a probability of 0.25. Our intent is to make a never-seen-before n-gram have a fairly, but not extremely, low conditional probability. The value 0.25 is somewhere between the 25^{th} percentile and the median $m_i(t)$ among all our training sets. For frequently occurring 4-grams, this modification leaves the value of m_i roughly unchanged. Jurafsky and Martin (2009) discuss many alternative smoothing algorithms for addressing this sparse data problem, but because of the relatively small size of our training corpus, many of these are infeasible.

We exclude the three months prior to month t from the training corpus because sometimes a 4-gram and its associated 3-gram, in the two year's prior to month t, may have occurred for the first time in month $t-1$. Furthermore if the associated 3-gram occurred as often in month $t-1$ as the 4-gram, the training set (unmodified) probability $P(w_4|w_1\ w_2\ w_3)$ will equal one, and the associated entropy contribution will be zero. However, this n-gram may still be "unusual" in month t if it has only been observed in month $t-1$ and at no other time in our training set. For example the 4-gram *a failed hedging strategy* is one of the top entropy contributors (see discusion in Section 3.3.1) in May 2012. It refers to the losses incurred in April and May of 2012 by the Chief Investment Office of JPMorgan. The 3-gram *a failed hedging* occurs for the first time in our sample in May 2012 as well, and both occur 53 times. Therefore, if May 2012 is included in the training corpus for June 2012, the conditional probability for this 4-gram will be one.[14] However, when this phrase appears (11 times) in June 2012, we would still like to regard it as unusual.

Our results are not very sensitive to any of these modeling assumptions (i.e. setting unob-

[13] We approximate $c(\{\omega_1\omega_2\omega_3\omega_4\})$ in a given training window by only counting the occurrences of those 4-grams which are among the most frequently occurring 5000 in every month. We therefore underestimate 4-gram counts, especially for less-frequently occurring n-grams, and therefore the m_i's associated with low p_i^{ν}'s are biased downwards. However, because $p \log p$ goes to zero for small p, this is unlikely to have a meaningful impact on our entropy measure. Across the 228 months in our sample, the maximum least-frequently-observed n-gram empirical probability is 0.012 percent. Rerunning the analysis using the top 4000 n-grams – instead of the top 5000 – in each month leaves our results largely unchanged, suggesting the analysis isn't sensitive to this issue.

[14] Using the modified $m_i(t)$ from (8) the probability would be 54/57.

served m_i's to 0.25, having the rolling window be two years, and the choice of three months for the training window offset).

3.3.1 Contribution to entropy

By sorting n-grams on their contribution to entropy in (6), we can identify for a given month the most and least unusual 4-word phrases. Table 4 shows the three top and bottom phrases[15] by their contribution to entropy in two months in our sample that had major market or geopolitical events: September 2008 (the Lehman bankruptcy) and May 2012 (around the peak of the European sovereign debt crisis). In each case, at least one of the n-grams with the largest entropy contribution reflects the key event of that month – and does so without any semantic context. On the other hand, the n-grams with the smallest entropy contribution are generic, and have no bearing on the event under consideration.

Consider for example the n-gram *nyse order imbalance _mn_* from September of 2008. In our training set, the majority of occurrences of the 3-gram *nyse order imbalance* were followed by *_n_* (a number) rather than *_mn_* (a number in the millions). The frequent occurrence of *nyse order imbalance* followed by a number in the millions, rather than a smaller number, is unusual. This 4-gram has a relatively large p_i, a low m_i (and a high $-\log m_i$), and is the top contributor to negative entropy in this month. On the other hand, the 3-gram *order imbalance _n_* is almost always followed by the word *shares*, thus giving this 4-gram an m_i of almost 1, and an entropy contribution close to zero. In May 2012, the n-gram *the euro zone crisis* is unusual because in the sample prior to this month the 3-gram *the euro zone* is frequently followed by *'s* or *debt*, but very infrequently by *crisis*. Therefore the relatively frequent occurrence in this month of this otherwise unusual phrase renders it a high negative entropy contributor.

While anecdotal, this evidence suggests that our entropy measure is able to sort phrases in a meaningful, and potentially important, way.

Aggregate entropy

We find that the aggregate entropy measures can be unduly influenced by a single frequently occurring n-gram. For example, if an n-gram i appears only in articles about one company in month t, but appears very often (i.e. has a large $p_i(t)$) and has a low model probability

[15] Some of the distinct 4-grams come from the same 5-gram.

$m_i(t)$, this one n-gram can distort the aggregate level entropy measure. A more stable measure of aggregate entropy is the first principal component of the single-name entropy series. For example, $ENTPOS$ can be measured as the first principal component of all the single-name $ENTPOS^j$ series. In the rest of the paper, all aggregate level entropy measures ($ENTALL(t)$, $ENTNEG(t)$, and $ENTPOS(t)$) are computed in this way.[16]

Figure 3 shows the three aggregate entropy series, with a scaled VIX superimposed. All three series are positively correlated with the VIX. $ENTPOS$ has the lowest correlation at 0.15, and $ENTNEG$ has the highest at 0.48. This is in contrast to the sentiment series where negative and positive sentiment have opposite signed VIX correlations. Since entropy reflects unusualness of news, it is perhaps not surprising that all entropy series are positively correlated with the VIX, as all news (neutral, positive, and negative) may be more unusual during times of high market volatility.

Table 3 shows the average single name (and non-principal component aggregate) entropy to VIX correlations. The average single names correlations for $ENTALL$ and $ENTNEG$ are positive, and the $ENTPOS$ average correlation is marginally negative though very close to zero. The values are smaller in magnitude than the correposnding sentiment ones. As espected, $SENTNEG$ ($SENTPOS$) is positively (negatively) correlated with contemporaneous implied volatility; both correlations are significantly different from zero.

The entropy series seem to reflect some of the same factors as the sentiment and VIX series, but also appear to have qualitatively different behavior. This gives hope that entropy contains information complementary to sentiment, which is the topic to which we now turn.

4 Single name volatility

At the single-name level, we explore the relationship between our news-based measures and future volatility in two steps. Section 4.1 shows that: our news-based measures contain relevant information about future volatility; that entropy and sentiment both matter; and that sentiment interacted with entropy contains more information than either measure on its own. Section 4.2 shows that our news-based measures remain useful forecasting tools even after we control

[16] Because of the need to have all data present for computing the principal component, our aggregate entropy measures use only 25 names for $ENTPOS$ and $ENTNEG$, and 31 names for ENT. For names that have observations at the start of sample period, but are missing some intermediate observations, we use the most recently available non-missing value of the associated entropy measure. See Footnote 20 for more details.

for known predictors of future volatility. Therefore, at the single-name level, the information contained in news-based measures is not already fully incorporated into prices.

4.1 Are news-based measures informative about future volatility?

We first want to establish that our entropy-based algorithm for extracting information from news stories does, in fact, contain useful information for future single-name volatility. We do not yet ask whether this information is already known to the market. We address this question in the next section.

To explore the extent to which our news-based measures contain information about future volatility, we regress single name implied volatility (30-day at-the-money) on lagged values of our news measures. The basic regression for name j has the form

$$IVOL^j_{1mo}(t+\phi) = a^j + b^{j\prime}\mathcal{L}_s\ NEWS^j(t) + \cdots + \epsilon^j(t) \tag{9}$$

where $NEWS^j$ is the news-based indicators under consideration, $\mathcal{L}_s$ is an s-lag operator, [17]and ϕ is an integer (set to zero for most of our results) which allows us to forecast volatility more than one month into the future. In our analysis s is set to either 3 or 6 months. The $\cdots$ in (9) indicates the possibility that additional right hand side variables will be present in the regression. We normalize all $NEWS$ variables to have unit standard deviation to make interpretation of coefficients easier. The j superscript usually indicates that the measure is computed from the list of n-grams coming from articles that mention company j.[18] We then average the time series regression coefficients across all names (for which we have implied volatility data) to obtain

$$b_l = \frac{\sum_j b^j_l}{\sum_j 1}. \tag{10}$$

We compute standard errors for each coefficient b_l in (10) under the assumption of independence.

To establish a baseline result for (9), we run the regression with $NEWS^j$ set to the percent of all month t articles that mention company j, which we call $ARTICLE\_PERCTOT^j$.[19] We use this measure because of our prior belief that it should contain minimal – though potentially

[17] $\mathcal{L}_s Y(t) = \{Y(t-1), Y(t-2), \ldots, Y(t-s)\}$.

[18] Only the article count measure doesn't look at n-grams (see $ARTICLE\_PERCTOT^j$ below).

[19] All results are qualitatively similar if we use the percent of all time t n-gram counts that come from articles that mention name j.

non-zero – information for future volatility. We refer to this as the *control* regression. Figure 5 plots the b coefficients, and associated confidence intervals. Indeed, all coefficients from (9) are not significantly different from zero.

The right chart in Figure (5) shows a plot of the fraction of all unadjusted R^2's of the single name regressions, using $ARTICLE\_PERCTOT$ on the right hand side, that are greater than a given value x, i.e. $f(x) = Pr(R^2 > x)$. Note that the x-axis in the chart starts at 1 and goes to 0. This function is one minus the cumulative distribution function of the R^2's from the single name regressions.

For an idealized zero-information regressor, this graph should be zero at all values of R^2 that are larger than 0, with a spike to probability one at $R^2 = 0$. The area under this curve would be zero. However, some single names have non-zero R^2's with respect to $ARTICLE\_PERCTOT^j$ and we have to control for this fact in interpreting the informativeness of our other regressors.

Similarly, for the ideal regressor with perfect explanatory power for every single name in our cross-section, the R^2 curve would spike up to 1 at $R^2 = 1$, and remain at 1 for all other potential R^2 values. The area under this curve would be 1. It is easy to show that the area under the $f(x)$ curve (AUC) is equal to the cross-sectional mean of R^2's from the single name regressions.

We will use the empirical $f(x)$ for $ARTICLE\_PERCTOT^j$ as the baseline R^2 curve (i.e. the one that obtains for a regressor with minimal predictive value). Comparing the AUC of the R^2 curves of other news-based variables to this one will tell us the incremental improvement in the cross-sectional average of R^2's that is achieved by a given regressor relative to a regressor with little predictive value. Furthermore, examining the shape of a given news-based R^2 curve relative to the baseline yields a richer picture of the predictive power of the measure in question than simply looking at the difference in cross-sectional means of R^2's.

Ranking news-based measures by their information content

The left two columns of Table 7 shows the difference in AUC's between our news-based measures and $ARTICLE\_PERCTOT$, or, equivalently, the difference in cross-sectional means of R^2's. We include the two sentiment indicators, the three entropy indicators, a variable that interacts negative sentiment with negative entropy ($ENTSENT\_NEG$), and another which interacts positive entropy with positive sentiment ($ENSENT\_POS$).[20] Results are shown for lags of 6

[20] In each single name regression, we exclude those months when one of the regressors is not available. For example, in a month where a given name had no n-grams classified as negative, while the negative sentiment

and 3 months in (9).

Consistent with some of the prior findings in the literature (for example, Tetlock (2007)) we find that negative sentiment contains information for future market outcomes – though Tetlock looks at stock returns and here we analyze implied volatility – offering an incremental improvement in average R^2's relative to the no-predictability benchmark of roughly 14 percent. Negative entropy yields an R^2 improvement of 9 percent. Positive sentiment and entropy do not contain incremental information.

Interestingly, the interacted variable, $ENTSENT\_NEG$, improves average R^2's by 22 percent, which is about double the improvement of either of the negative news measures separately. Figure 6 shows the results of this regression. The difference in the R^2 curve relative to the no-predictability benchmark is dramatic. All 6 lagged coefficient are statistically significant. The coefficient estimates are economically very large. A one unit standard deviation increase in last month's $ENTSENT\_NEG$ will increase this month's one month implied volatitility by 4 volatility points on average. Though, it should be emphasized, this information may already be incorporated in prices.

As a robustness check, we run the regression in (9) with $SENTNEG$, $ENTNEG$, and $ENTSENT\_NEG$ as the regressors. The control regression uses $ARTICLE\_PERCTOT$, $NGRAM\_PERCTOT$ and $ARTICLE\_PERCTOT \times NGRAM\_PERCTOT$. As Figure 7 shows, when all three variables are included, negative sentiment and entropy are statistically and economically marginal, while the interacted term $ENTSENT\_NEG$ remains both statistically significant and economically large.

Past work has used sentiment as *the* measure of the information content of news (see Tetlock (2007), Tetlock et al. (2008), Garcia (2013), Jegadeesh and Wu (2014), among others). Our results show that sentiment interacted with unusualness contains significantly more information about future implied volatility, at the single name level, than either sentiment or entropy on its own. As will become apparent, this finding holds in most of the other results in this paper.

measure is zero, the negative entropy measure from (6) is not defined. Replacing all such unobservable entropy scores with zero slightly reduces the magnitude of our results, but does not change any of the qualitative conclusions.

4.2 Is this information already in the price?

An important question is whether the information present in our news-based measures is already known to the market. Given that our sample contains 50 of the largest – and therefore most closely followed by investors – financial firms in the world, and that our analysis is at a monthly time horizon, the bar for finding information in our news-based measures that is new to market participants is quite high.

Bekaert and Hoerova (2014) show that, at the index level, lags of implied variance, realized variance, and stock price jumps all matter for forecasting future realized variance. To control for these effects, we use our 30-day at-the-money implied volatility measure $IVOL$, 20 trading-day realized volatility $RVOL$, and the negative and positive portions of monthly returns r^+ and r^- as explanatory variables for future realized and implied volatility.[21] Our basic specification for evaluating the forecasting power of a news-based measure $NEWS^j$ is the following panel regression:

$$\begin{aligned} VOL^j(t) = a^j &+ c_1' \mathcal{L}_s RVOL^j_{30day}(t) + c_2' \mathcal{L}_s IVOL^j_{1mo}(t) + c_3' \mathcal{L}_s r^{-j}(t) \\ &+ b1' \mathcal{L}_s ARTICLE\_PERCTOT^j(t) + b2' \mathcal{L}_s NEWS^j(t) + \epsilon^j(t), \end{aligned} \tag{11}$$

where VOL is either either $IVOL$ or $RVOL$, and a^j is an individual fixed effect term. The variable $ARTICLE\_PERCTOT^j$ is intended to control for the information content of news volume. As in the single name regressions, all news measures are normalized to have unit variance.

We show results for $s = 2$ (the ones for $s = 3$ are qualitatively similar and aren't shown to conserve space). We run this specification in variance, log variance and volatility terms. All of these yield similar qualitative results. We show the volatility results in the paper because these are the easiest to interpret. Also, adding r^+ as an explanatory variable was not impactful in any of our specifications, so we do not include this variable in our regression results.

Before turning to the forecasting regression in (11), we examine briefly the drivers of our news-based measures. The following is our descriptive panel specification:

$$NEWS^j(t) = a^j + c_1' \mathcal{L}_2 RVOL^j_{30day}(t) + c_2' \mathcal{L}_2 IVOL^j_{1mo}(t) + c_3' \mathcal{L}_2 r^{-j}(t) + b' \mathcal{L}_2 NEWS^j(t) + \epsilon^j(t). \tag{12}$$

[21] $r^- \equiv \max(-r, 0)$ and $r^+ \equiv \max(r, 0)$.

This is run for each of the following categories of news measures:

- *positive:* $SENTPOS$, $ENTPOS$, $ENTSENT\_POS$;
- *negative:* $SENTNEG$, $ENTNEG$, $ENTSENT\_NEG$.

Table 8 shows the results of this descriptive regression. While lagged volatility has little effect on the positive sentiment news measures, high past realized volatility forecasts higher negative sentiment news measures in the future. Absence of past negative returns forecasts higher future positive news measures, whereas the presence of negative returns forecasts higher future negative news measures. The positive and negative news measures are less persistent than percent article counts – thus news sentiment about a firm quickly reverts back to its average levels, whereas a firm may stay "in the news" for relatively longer.

Tables 9 and 10 show the results of the specification in (11) for implied and realized volatility respectively. The control variables (lagged $IVOL$, $RVOL$, and r^-) all matter for both future realized and implied volatility, and enter the panel with the expected positive sign (only $r^-(t-2)$ enters with a negative sign, and that just partly offsets the large positive loading on $r^-(t-1)$).

The positive category news measures (Models 1–3) all show up with negative coefficients, suggesting positive news at time $t-1$ or $t-2$ forecast lower time t implied and realized volatility, after controlling for known forecasting variables. Adding the lag 1 and lag 2 coefficients on $NEWS^j$, reported in the row labeled "Sum Last Two", allows us to evaluate the importance of the difference news measures. We find that $ENTSENT\_POS$ has a larger effect on future volatility than either positive sentiment or entropy on their own. Furthermore the economic significance of the effect is large. For example, as Table 9 shows, these two coefficients are -1.04 for $ENTSENT\_POS$, suggesting that a one standard deviation increase in current positive and unusual news forecasts a 1 volatility point drop (e.g. from 20 to 19) next month. The results for future realized volatility in Table 10 are similar.

The negative category news measures (Models 4–6) forecast future implied and realized volatility with a positive sign. All three news-based measures ($SENTNEG$, $ENTNEG$, $ENTSENT\_NEG$) are economically and statistically meaningful, with the interacted term $ENTSENT\_NEG$ having the largest economic effect. A one standard deviation increase in the latter implies a 1.546 (2.428) rise in next month's implied (realized) volatility.

Model (7) which has article percent count as the sole news-based measure offers some evidence that firms that are in the news a lot, irrespective of sentiment, tend to have lower implied and realized volatilities in future months.

Our panel results suggest that, even after controlling for known determinants of future volatility, our news-based measures still contain useful forecasting information. The coefficient estimates on lagged news-measures were shown to be statistically and economically meaningful. Furthermore, for both the positive and negative sentiment categories, the interacted news terms ($ENTSENT\_POS$ and $ENTSENT\_NEG$) contain more information than either sentiment or entropy on their own. This is a recurrent theme in this paper's results.

A robustness check using R^2's

The third and fourth columns of Table 7 show the results of the single-name regressions in (9) when lags of implied volatility are included as explanatory variables. The analysis is identical to the information rankings of the news-based measures that were discussued in Section 4.1, except we now add lags of implied volatility to both the control and evaluation regressions. Adding lagged implied volatility increases the average R^2 of the control regressions from 0.105 (0.0796) with 6 (3) lags to 0.69 (0.549). Thus, lagged implied volatility contains a great deal of information for future implied volatility. The incremental contribution of our news-based measures to the difference of average R^2's between the control and evaluation regressions drops meaningfully from the specification which did not include lagged implied volatility. While different news-based measures add some incremental explanatory power as measured by average R^2 (especially in the $\phi = 1$ specification), the size of the effect is economically small.

These results suggest that much of the explanatory power we found in Section 4.1 was due to the contemporaneous correlation of our news-based measures with implied volatility. However, the results from the panel specification in (11) show that our news-based measures contain significant and economically meaningful content for future volatility after controlling for known forecasting variables. The discrepancy between these two sets of results suggests that either the panel regression has more power to reject the null of no incremental information in the news-based measures, or that while significant, the economic impact of the news-based measures for forecasting future volatility is small for the set of names we consider. A deeper examination of this discrepancy using a dataset with more and smaller firms is an interesting area for future work.

5 Aggregate volatility

We now turn from company-specific measures of entropy, sentiment, and volatility to aggregate measures. We document evidence that unusual negative news predicts an increase in volatility as measured either by the VIX or by realized volatility on the S&P 500 index. As discussed in Section 3, each aggregate measure of entropy is the first principal component of the corresponding measures across the financial companies listed in Table 1, whereas aggregate sentiment follows from (7) applied to the set of all n-grams in month t.

We consider the five aggregate news-based measures from Figures 2 and 3, as well as the interaction variable $ENTSENT\_NEG(t) = ENTNEG(t) \times SENTNEG(t)$. Table 5 gives some descriptive statistics about these measures, and Table 6 shows the contemporaneous correlations among these six variables, and the VIX index. Figure 4 shows a plot of $ENTSENT\_NEG$ versus the VIX index.

$SENTPOS$ has a negative correlation with the VIX, whereas all the entropy measures have a positive correlation, suggesting that at the aggregate level, news unusualness increases with market volatility. All entropy measures are positively correlated with one another, and negatively correlated with $SENTPOS$.

It is notable that although $ENTNEG$ and $SENTNEG$ have a low correlation of 0.19, their correlations with the VIX are 0.48 and 0.46 respectively. So even though the two do not share much in common, it appears they both explain a meaningful portion of VIX variability. The interaction variable $ENTSENT\_NEG$ has the highest VIX correlation of the news based measures at 0.6. It also has a high correlation with its consituents: 0.86 with $SENTNEG$ and 0.64 with $ENTNEG$.

This correlation result, the visual evidence in Figure 4 and the desriptive statistics in Table 5 all suggest that the interacted variable $ENTSENT\_NEG$ is a closer fit to the VIX (and realized volatility) than either negative sentiment or entropy separately.

In the next two sections, we explore the dynamics of this relationship in greater detail.

5.1 Event studies

For a first look at the data, we examine changes in the VIX around high and low values of our aggregate measures. For each aggregate measure (such as $ENTNEG$ or $SENTNEG$), we sort the 177 months from April 1999 through December 2013 according to the value of the measure and select the months in the top and bottom and quintiles. We think of the months in these quintiles as event dates. For each such month, we record the level of the VIX over the 25-month period starting 12 months before the event and ending 12 months after. We then average the VIX paths across the months in each quintile to see the average behavior of the VIX around one of these events.

As a point of reference, Figure 8 shows the results when the events are high and low values of the VIX itself. The dashed lines are two standard errors above and below the solid average line. As expected, the left panel shows that the VIX increases to a peak and then declines; the right panel confirms that the VIX decreases and then increases around a low value, but the pattern is much less pronounced around a low point than around a high point. In part for this reason, we focus primarily on the quintile associated with high volatility when we sort on other variables.

Figure 9 shows corresponding event studies for various measures, starting with $ENTNEG$ in the first row. Around a high level of $ENTNEG$, we see the VIX first climbing and then staying elevated, in contrast to the sharp mean-reversion we see in Figure 8. Around a low level of $ENTNEG$, the drop and rebound in the VIX is more pronounced than it is around a low level of the VIX itself in Figure 8. High levels of $SENTNEG$ have a similar association with the VIX, but low levels of $SENTNEG$ are associated with a steady decline in the VIX, unlike the pattern around low levels of $ENTNEG$. Around a high level of the interaction variable $ENTSENT\_NEG$ we again see a climb in the VIX but almost no subsequent decline — a high level of the $ENTSENT\_NEG$ variable signals a sustained elevation in volatility.[22] We interpret this as further evidence that unusual negative news forecasts market stress. Indeed, the effect is large, with a high level of $ENTSENT\_NEG$ associated with VIX increase of almost 10 points. The effect lasts for months, consistent with the findings in the company-specific regressions of Section 4. As a final point of comparison, the lower-right panel of Figure 8 shows that high levels of overall entropy have no association with changes in the VIX.

We obtain qualitatively similar results using realized volatility (measured as the standard deviation of daily returns within a month) instead of the VIX. For example, the left panel of Figure 10 shows the evolution of realized volatility around top quintile levels of $ENTSENT\_NEG$: as with the VIX, realized volatility climbs to month zero and remains elevated, declining only slowly after the peak. The right panel of Figure 10 shows the behavior of the volatility risk premium, measured as the difference between the VIX and realized volatility. The

[22]This behavior for the interacted variable does not automatically follow from similar behavior for $ENTNEG$ and $SENTNEG$ because high levels of these variables need not occur together.

volatility risk premium declines slightly, indicating that realized volatility increases a bit more than implied volatility around high levels of $ENTSENT\_NEG$ and suggesting that elevated $ENTSENT\_NEG$ is associated with increased market stress and not simply increased risk aversion. The figure uses end-of-month VIX values, but the pattern remains the same if we use beginning-of-month VIX values to calculate the volatility risk premium.

5.2 Impulse Response Functions

We next investigate interactions among the aggregate variables through vector autoregressions (VARs). The event studies of the previous section have the advantage of being nonparametric. A VAR model imposes more assumptions but also provides a more systematic analysis, so the perspectives complement each other.

We estimate a VAR model in six variables, initially ordered as follows: VIX, SPX_RVOL (realized volatility), $SENTNEG$, $ENTSENT\_NEG$, $SENTPOS$, and $ENTSENT\_POS$.[23] The Akaike information criterion selects a model with two lags; we estimate each equation in the VAR separately using ordinary least squares. We analyze the model through its impulse response functions. Each impulse is a one standard deviation shock to the error term for one variable in a Cholesky factorization of the error covariance matrix. A shock to one variable has a direct effect on variables listed later in the order of variables but not on variables listed earlier. Our ordering is thus stacked against finding an influence on either measure of volatility from the entropy and sentiment measures.

The left panel of Figure 11 shows impulse response functions in response to a shock to $ENTSENT\_NEG$, together with bootstrapped 95 percent confidence intervals.[24] Both the VIX and realized volatility have statistically significant responses to the shock. A one standard deviation increase in $ENTSENT\_NEG$ increases the VIX by 1.5 points and increases realized volatility by two points, so a two to three standard deviation shock to $ENTSENT\_NEG$ has a large economic impact on volatility. The right panel shows corresponding results in response to a shock to $SENTNEG$. There, neither VIX nor realized volatility exhibits a statistically significant response.

Next we reverse the order of $ENTSENT\_NEG$ and $SENTNEG$ and recalculate the impulse

[23] Running the analysis in variance or log variance terms, with or without r^- as one of the model variables, does not change any of our results. We focus on the volatility model that excludes r^- for simplicity.

[24] We used the **R** package "vars" for the VAR estimation and impulse response functions; see Pfaff (2008).

response functions. The left panel of Figure 12 shows that the VIX and realized volatility now have statistically significant responses to $SENTNEG$, increasing by roughly 1.25 and 1.75 points, respectively. But the right panel shows that they still have marginally significant responses to $ENTSENT\_NEG$ following the order change. Taking Figures 11 and 12 together suggests the following conclusions: An increase in negative sentiment or its interaction with entropy each predicts an increase in volatility; the effect of negative sentiment is captured by the interaction term; but there is an effect from the interaction term that is not captured by negative sentiment alone. This is consistent with our findings in the company-specific regressions of Section 4.

Figures 13 and 14 show that a similar pattern holds for positive sentiment and its interaction with entropy. A shock to the interaction variable $ENTSENT\_POS$ has a statistically significant (negative) effect on both VIX and realized volatility when it is listed before $SENTPOS$ (Figure 13, left panel), whereas $SENTPOS$ does not (Figure 13, right panel). When the order of the variables is interchanged, $SENTPOS$ has a statistically significant effect on VIX (Figure 14, left panel), and $ENTSENT\_POS$ has a marginally significant effect on both VIX and realized volatility (Figure 14, right panel). As one would expect the magnitudes of the responses to the positive signals are smaller than the responses to the negative signals, but the overall pattern is similar. The pattern suggests that both positive sentiment and its interaction with entropy influence volatility, and that the interaction term captures an effect that is not present in the sentiment variable alone.

The time horizon of the impulse responses is also noteworthy. Consider, for example, the two responses in the upper left portion of Figure 11. They show that the effect on volatility of an increase in $ENTSENT\_NEG$ plays out over months, peaking around four months after the shock and dissipating slowly. In Section 4 we found that the corresponding coefficients in the company-specific regressions remain statistically significant at lags of several months. These time scales are markedly different from those in prior work using news sentiment to predict returns (including Da et al. 2014, Jegadeesh and Wu 2014, Tetlock 2007, and Tetlock et al. 2008), where effects play out over days. In other words, directional information is incorporated into prices within days, but signals forecasting elevated volatility can remain relevant for months.

Volatility is of course much more persistent than returns are, but this property is insufficient to explain the volatility responses in Figures 11–14. Including implied and realized volatility in the VARs controls for persistence. Although persistence of volatility could make a predictor of high volatility in the present a predictor of high volatility in the future, the impulse responses of

VIX and realized volatility to the news variables are consistently hump-shaped wherever they are statistically significant. The responses at month four are therefore not simply lingering effects of a larger response in month one, as persistence by itself would predict. In Section 7, we will see that the hump-shaped responses are consistent with a simple model of rational inattention of agents who face constraints on the volume of information they can process.

6 Return predictability

Much of prior work on textual analysis in finance has focused on predicting returns. While the focus of our work has been on forecasting market stress, we want to briefly investigate whether our "unusualness"-based news measures are useful for predicting returns, to place our work into the context of the broader literature. It should be noted that our sample of companies is small (only 50 firms), and all companies are in related industries (finance, insurance, real estate). This lack of company and industry diversification stacks the cards against finding evidence of return predictability. Furthermore, any results we do find may be unduly influenced by outliers in our small sample of firms. Consequently, the results in this section are only indicative, and should be interpreted with caution.

We form long-short portfolios of single names in month t based on two different news-based sorts. We choose some fraction (4 percent, 10 percent or 20 percent) of companies for each of the long and short portfolios, and then hold each company chosen based on the month t signal for either one, three, or six months. The portfolio in month t holds an equal weight in stocks identified over the prior one, three, or six months. When we do not have data in a given month for a company that is part of the long or short portfolio, that name is exluded from that month's return calculation. We use U.S. dollar total returns for all names in our sample, and approximate month t dividends by the last twelve month dividend yield divided by 12.

The two news-based sorts for forming portfolios are:

- $SENTNEG$: Shorts (longs) are some fraction of names j with the highest (lowest) month t values of $SENTNEG^j$.
- $SENTPOS$ vs $SENTNEG$: Shorts (longs) are some fraction of names with the highest values of $SENTNEG^j$ ($SENTPOS^j$) in month t. This is similar to the portfolio scheme from Tetlock, Saar-Tsechansky, and Macskassy (2008).

For each of these sorts, we compute a competing sort where the sentiment measure is interacted with $ENTALL^j$ – we refer to this as the *interacted* sort. Ideally, we would like to interact positive and negative sentiment with positive and negative entropy, as we've done elsewhere in the paper. However, as mentioned in Footnote 20, time t negative and positive entropies are frequently not available, and extrapolating from past entropy introduces too much noise. So to maximize the number of company-months of interacted news measures, we use $ENTALL^j$ – which is almost always available for all names – as the interacting variable.

Figure 15 shows the cumulative returns for two sets of portfolios: negative sentiment with a six month holding period where the fraction of companies chosen in a given month is 4 percent; and positive vs negative sentiment for a one month holding period with a portfolio fraction of 20 percent. The data starts from April 1998 (the first month for which we calculate news-based measures in our sample) and ends in December 2014. Note that the extremely high returns in November and December of 2014 for the negative sentiment sort long-short portfolios are correct, and are due to the portfolio's being long Chinese banking stocks, which had large rallies in those two months. In both cases, the interacted sort outperforms, with the economic effect being especially large in the case of $SENTPOS$ vs $SENTNEG$.

To gain further insight into these results, we regress the monthly returns of the long-short portfolios from Figure 15 on the Fama-French global factors (market, size, value) and on the global momentum factor.[25] Table 11 shows the results. First we see that all news-based portfolios have very little overlap with any of the Fama-French factors, with the R^2's of all regressions being effectively zero. The portfolios exhibit a small positive loading on the small minus big (SMB) factor and on the high minus low (HML) factor. Our portfolios thus somewhat mimick small, value stocks (even though all stocks in our sample are large). While the alphas are economically large (around 50 basis points per month), our tests have low power as the only significant alpha is from the interacted negative-positive sentiment sort. The interacted monthly alphas are roughly 20-30 basis points higher than the non-interacted ones. Unusualness interacted with sentiment outperforms sentiment only sorts in this example.

As a robustness check, Table 12 (for the negative sentiment portfolios) and Table 13 (for the positive versus negative sentiment portfolios) report annualized Sharpe ratios and monthly alphas of all fraction/holding-period combinations that we analyzed. Each cell in the table reports, for a given fraction and holding period pair, the Sharpe ratio and four factor alpha of the sentiment portfolio, the Sharpe ratio and four factor alpha of the sentiment portfolio

[25] Data on Fama-French global factors are obtained from Ken French's website using the Quandl API for **R**.

interacted with entropy, and the difference of the Sharpe ratio and alpha. The pattern that the entropy interacted portfolios outperform the sentiment-only portfolios is robust across most of the combinations of parameters that we examined.

While the results of this section should be interpreted with caution, we again find evidence suggesting that (1) news-based measures contain information useful for forecasting future market outcomes that is not already known to market participants, and (2) sentiment interacted with entropy is more informative than sentiment on its own.

7 Rational inattention and information constraints

Several studies have found evidence that the limits of human attention affect market prices. Dellavigna and Pollett (2009) find a less immediate response to earnings announced on Fridays than other days and explain the differences through reduced investor attention. Ehrmann and Jansen (2012) document changes in the comovement of international stock prices during World Cup soccer matches, when traders are presumably distracted. Huberman and Regev (2001) and Tetlock (2011) document striking stock market responses to "news" that was previously made public. Hirshleifer, Hou, Teoh, and Zhang (2004) explain stock return predictability from accounting data through limited investor attention. Corwin and Coughenour (2008) find that attention allocation by market specialists affects transaction costs. Sicherman et al. (2015) document patterns of investor attention in response to market conditions. Daniel, Hirshleifer, and Teoh (2002) explain a broad range psychological effects on markets through limited attention.

Limited attention may help explain the patterns we observe in Sections 4 and 5. Searching news articles to extract information about unusualness and sentiment takes time, and investors may perceive that they have better options for gathering data with whatever resources they allocate to making investment decisions. Consistent with Samuelson's dictum (Jung and Shiller 2005), investors may focus on a small set of stocks and pay less attention to macro events.[26] In Dellavigna and Pollett (2007), investors focus on information relevant to near-term returns but are inattentive to information with long-term consequences. A similar effect may apply in our setting, albeit over a shorter horizon. This would be consistent with the impulse response functions for volatility in Section 5.2, in which the response to a signal is greater at intermediate

[26] In the model of Peng and Xiong (2006), investors choose instead to focus on coarser aggregate information and pay less attention to idiosyncratic information. For our purposes, the point is that this is one of the dimensions along which agents need to make an attention allocation decision.

horizons than at the shortest horizons.

We can develop a stronger connection between the impulse response functions and limited attention by building on work of Sims (2003, 2015) and Maćkowiak and Wiederholt (2009ab). Sims (2015) presents a theoretical framework, developed in a series of papers starting with Sims (2003), for modeling rational inattention.[27] Agents face constraints or costs on information processing and incorporate these into rational choices. Maćkowiak and Wiederholt (2009ab) build on Sims's framework to model sticky prices for goods; in their setting, a firm allocates limited attention capacity to two types of information, aggregate and idiosyncratic. The qualitative implications of reduced attention are relevant to our setting.

To develop the connection, we will let X_t denote a macro state variable such as the reciprocal of aggregate consumption or its negative logarithm.[28] For simplicity, we suppose that X_t follows a stationary AR(1) process,

$$X_{t+1} = \rho X_t + a u_{t+1}, \tag{13}$$

where $\rho \in (0, 1)$, and the $\{u_t\}$ are independent, standard normal random errors.

Agents would like to hedge macro risk associated with X_t. However, they face information constraints that prevent them from observing X_t precisely; these constraints reflect intrinsic difficulty in measuring the macro state as well as the limits of agents' attention capacity. As a consequence, agents cannot write contracts with payoffs directly determined by X_t. Instead, they write contracts on an approximation Y_t that solves

$$\min_{b,c} E[(X_t - Y_t)^2]$$

with

$$Y_t = \sum_{\ell=0}^{\infty} b_\ell u_{t-\ell} + \sum_{\ell=0}^{\infty} c_\ell \epsilon_{t-\ell}, \tag{14}$$

subject to an information constraint between the processes $\{Y_t\}$ and $\{X_t\}$. The $\{\epsilon_t\}$ form a sequence of independent, standard normal random errors independent of $\{u_t\}$. Interpret Y_t as the best approximation to the macro state X_t given the information constraint.[29]

[27] Sims (2003), p.696, makes an explicit connection with the saliency of information in news media.

[28] This formulation makes agents averse to large values of X_t and will simplify the interpretation of the VIX as a hedge for macro risk.

[29] We omit the precise definition of the information constraint because it takes several steps to develop. In the

Maćkowiak and Wiederholt (2009a) show that the effect of the information constraint is equivalent to having agents observe a noisy signal $S^t = \{\ldots, S_0, S_1, \ldots, S_t\}$ of the past rather than the complete history $\{\ldots,\ (u_0, \epsilon_0),\ (u_1, \epsilon_1),\ \ldots,\ (u_t, \epsilon_t)\}$. In particular, $Y_t = E[X_t|S^t]$, meaning that the best observable approximation to the macro state is the conditional expectation of the true macro state given the agents' available information.

Next we consider the price at time t of a contract paying Y_{t+1} at time t. We assume a stochastic discount factor of the form[30] $\exp(\lambda u_{t+1} - \lambda^2/2)$, where u_{t+1} is the innovation to the macro state in (13). This factor attaches a greater discount to cash flows that covary negatively with shocks to X_t. Ordinarily, the price at time t would be the time-t conditional expectation of the product of the payoff and the stochastic discount factor. Given agents' limited information S^t about the past, we model the price as[31]

$$V_t = E\left[e^{\lambda u_{t+1} - \lambda^2/2} Y_{t+1} | S^t\right].$$

The key implication of this formulation (derived in the appendix) is that the impulse response of V_t to a shock to the error in X_t is hump-shaped if agents' information processing constraint is sufficiently tight.

To map this observation to the impulse response functions in Section 5.2, think of the VIX as the price of a contract that imperfectly hedges macro risk: higher levels of the VIX are associated with market stress, so a contract that pays more in these states partly offsets a macro risk. Interpret the aggregate variable $ENTSENT\_NEG$ as a proxy for the macro state. The model we have outlined predicts that when the information constraint between $ENTSENT\_NEG$ and the VIX is tight, the impulse response function should be humped, just as we saw in Section 5.2. The information constraint faced by agents limits how quickly innovations to the macro state get incorporated in the VIX.

A more precise mapping between the model and our application should recognize that

case of scalar (jointly) normal random variables, the constraint reduces to an upper bound on their correlation. The general definition is detailed in Maćkowiak and Wiederholt (2009ab), and relevant background from information theory is reviewed in Sims (2003, 2015). The resulting Y_t is optimal among approximations with the moving average representation in (14).

[30] We assume an interest rate of zero for simplicity. If we interpret negative X_t as the log of aggregate consumption, then for a representative agent with power utility, a time discount coefficient δ and a risk-aversion parameter λ, the stochastic discount factor would be $\exp(-\delta + \lambda(au_{t+1} - (1-\rho)X_t))$. In this case, the one-period interest rate is $r_t = \delta + \lambda(1-\rho)Y_t - c$ for some constant $c \geq 0$ and $Y_t = E[X_t|S^t]$. In our discussion, we use a simplified version of the stochastic discount factor for clarity of exposition.

[31] Peng and Xiong (2006) develop a theoretical framework for asset pricing in a model with rational inattention. Our pricing formula has the same general structure as their equation (71).

$ENTSENT\_NEG$ is itself at best a noisy observation of the macro state, say $ENTSENT(t) = X_t + \sigma_\eta \eta_t$, for some independent error term η_t. Then a one standard deviation shock to $ENTSENT\_NEG$ combines shocks to u_{t+1} and η_{t+1}, but $\{V_t\}$ responds only to the shock to u_{t+1}. In this interpretation, the impulse response functions we observe in Section 5.2 are averages over the responses to random shocks u_{t+1} in the unseen X_t, conditional on the total shock to the error in $ENTSENT\_NEG$ equaling one standard deviation. The average impulse response preserves the hump shape at least if the error variance σ_η^2 is not too large.

8 Conclusion

Using techniques from natural language processing, we develop a methodology for classifying the degree of "unusualness" of news. Applying our measure of unusualness to a large news dataset that we obtain from Thomson Reuters, we show that unusual negative news forecasts volatility at both the company-specific and aggregate level. News shocks are impounded into volatility over the course of several months. This is a much longer time horizon than previous studies – which have focused on returns rather than volatility – have documented.

Across multiple analyses, we find that interacted measures of unusualness and sentiment provide the best predictors of future volatility among the news measures we study:

- In company-specific regressions of implied volatility on lagged news measures, unusual negative news is economically and statistically significant at lags of up to six months. It increases average R^2 across companies by more than sentiment or unusualness measures alone.

- Our interacted measure remains significant when we control for other predictors of volatility (lagged volatility measures and negative returns), indicating that the information in this news measure is not fully reflected in contemporaneous prices.

- At the aggregate level, we run vector autoregressions of the VIX and realized market volatility with several aggregate news variables. Impulse response functions show that a shock to our interacted measure of unusual negative news predicts an increase in implied and realized volatility over several months. The effect is stronger for our interacted variable than for negative sentiment alone.

- We measure the performance of long-short portfolios sorted on sentiment measures and on sentiment measures interacted with unusualness measures. Across multiple portfolio rules, sorts based on the interacted measures typically outperform sorts based on sentiment alone.

In our aggregate analysis, we find that news shocks affect realized and implied volatilities in a hump-shaped manner over time. This response would not obtain simply from the persistence of volatility: in this case the effect of a news shock would dissipate monotonically. A hump-shaped response indicates that news is not absorbed by the market instantaneously. We argue that this type of response is consistent with investors who face constraints on the rate at which they can process information.

Using tools from the rational inattention literature, we develop a simple model of the price of a security which tracks the true macro state of the world subject to an informational flow constraint. When the flow rate is sufficiently restricted, the model generates a hump-shaped price response to macro innovations.

The connection we make between this market friction and an empirical measurement of how market prices incorporate news is novel, and leads to many interesting research questions. Primary among these is how to relate our results to Samuelson's dictum on micro- vs macro-efficiency. We hope to pursue this question in future research.

Finally, because of our finding that news is incorporated into market volatility only gradually, our methodology should prove useful for risk monitoring.

A Appendix

A.1 Data cleaning

This section summarizes our data cleaning methodology. Further details are available from the authors.

Articles whose headlines begin with `REG-` (regulatory filings) and `TABLE-` (data tables) are deleted. The `reuters` tag at the start of an article and in the end-of-article disclaimer is removed, as is any additional post article information identifying the author of the article.

Punctuation characters (`,` or `;` and so on) and quotation marks are deleted, as are prefixes and suffixes that are followed by a period (e.g. `mr`, `corp`, etc.). All known references to any of the fifty companies in our sample are replaced with the string `_company_`.[32] Different references to the same, multi-word entity are replaced with a unique string. For example, all variations of `standard & poor's` are replaced with `snp`, references to `new york stock exchange` are replaced with `nyse`, and so on.

References to years, of the form `19xx-xx` or `20xx-xx` or similar forms, are replaced with `_y_`. We replace all numbers identified as being in the millions (billions) with `_mn_` (`_bn_`). Other numbers or fractions are replaced with `_n_`. The symbols `&` and `$` are deleted. All references to percent (e.g. `%` or `pct` or `pctage` etc.) are replaced with `pct`.

We make an attempt to delete all references to email addresses or web sites, though we do not have a systemic way of doing so.

Following this text processing step, we use the NLTK package from Python to convert the raw text into n-grams. First `sent_tokenize()` segments the text into sentences. Then `word_tokenize()` breaks the sentence into single words. In this step, standard contractions are split (e.g. `don't` becomes `do` and `n't`). Finally `ngrams()` is used to create 3- and 4-grams from the post-processed, tokenized text.

A.2 Rational inattention

Proposition 3 of Maćkowiak and Wiederholt (2009a) shows that the optimal Y_t in (14) has

$$b_\ell = a\left(\rho^\ell - \frac{1}{2^{2\kappa}}\left(\frac{\rho}{2^{2\kappa}}\right)^\ell\right), \tag{15}$$

and

$$c_\ell = c_0\left(\frac{\rho}{2^{2\kappa}}\right)^\ell, \tag{16}$$

where κ is the upper bound constraint on the information flow rate between the sequences $\{X_t\}$ and $\{Y_t\}$; see also Section 3.2.2 of Sims (2015). The definition of the information flow rate is detailed in Maćkowiak and Wiederholt (2009ab), and relevant background from information theory is reviewed in Sims (2003, 2015). At $\kappa = \infty$, $b_\ell = a\rho^\ell$ and $c_\ell = 0$, so Y_t coincides with the moving-average representation of the AR(1) process X_t. At $\kappa = 0$, we have $b_\ell = 0$, and no information about $\{u_t\}$ is incorporated into Y_t; in fact, Y_t is identically zero in that case because

[32] It is likely that we have not identified all possible references to companies in our sample.

$c_0 = 0$ at $\kappa = 0$.

The innovation u_{t+1} is independent of past values of u_t and ϵ_t, and it remains so conditional on the agents' information S^t. A standard calculation for normal random variables therefore gives

$$E\left[e^{\lambda u_{t+1} - \lambda^2/2} Y_{t+1} | S^t\right] = E[b_0 \lambda + Y_{t+1} | S^t].$$

It follows from (15)–(16) (and is shown explicitly in Appendix G of Maćkowiak and Wiederholt 2009a) that

$$Y_{t+1} = \left(\frac{\rho}{2^{2\kappa}}\right) Y_t + \left(1 - \frac{1}{2^{2\kappa}}\right) X_{t+1} + c_0 \epsilon_{t+1}.$$

Replacing X_{t+1} with the right side of (13) and using the fact that $E[X_t | S^t] = Y_t$ (proved in Appendix H of Maćkowiak and Wiederholt 2009a) we get

$$E[Y_{t+1} | S^t] = \left(\frac{\rho}{2^{2\kappa}}\right) Y_t + \left(1 - \frac{1}{2^{2\kappa}}\right) \rho E[X_t | S^t] = \rho Y_t$$

and then

$$V_t = b_0 \lambda + \rho Y_t.$$

The price premium $b_0 \lambda$ increases with κ because b_0 does. In other words, the contract is worth more with looser information constraints because it yields a better hedge in that case.

Given this representation and (15), the response of $V_t, V_{t+1}, \ldots$ to an impulse of $u_t = 1$ is given by $b_0 \lambda + \rho b_t$, $t = 0, 1, \ldots$. As illustrated in Figure 16, for small values of κ, this is a hump-shaped function of t, and for large values of κ it decreases monotonically.

References

Baker, S., N. Bloom, and S. Davis, 2013, "Measuring economic policy uncertainty," working paper.

Bekaert, G. and M. Hoerova, 2014, "The VIX, the variance premium and stock market volatility," *Journal of Econometrics*, 183 (2), 181–192.

Bisias, D., M. D. Flood, A. W. Lo, and S. Valavanis, 2012, "A survey of systemic risk analytics," Working paper 1, U.S. Department of Treasury, Office of Financial Research.

Corwin, S., and J. Coughenour, 2008, "Limited attention and the allocation of effort in securities trading," *Journal of Finance*, 63(6): 3031–3067

Da, Z., J. Engelberg, and P. Gao, 2014, "The sum of all FEARS invesor sentiment and asset prices," *The Review of Financial Studies*, 28 (1), 1–32.

Daniel, K., D. Hirshleifer, and S. H. Teoh, 2002, "Investor psychology in capital markets: evidence and policy implications," *Journal of Monetary Economics*, 49, 139–209.

Dellavigna, S., and J. M. Pollet, 2007, "Demographics and industry returns," *American Economic Review*, 1667–1702.

Dellavigna, S., and J. M. Pollet, 2009, "Investor inattention and Friday earnings announcements," *Journal of Finance*, 64, 709–749.

Ehrmann, M., and D.-J. Jansen, 2012, "The pitch rather than the pit: investor inattention during FIFA World Cup matches," Working Paper 1424, European Central Bank.

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

Hendershott, T., D. Livdan, and N. Schürhoff, 2014, "Are institutions informed about news?" working paper.

Heston, S. and N. Sinha, 2014, "News versus sentiment: Comparing textual processing approaches for predicting stock returns," *working paper.*

Hirshleifer, D., 2001, "Investor psychology and asset pricing," *Journal of Finance*, 56, 1533–1597.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang, 2004, "Do investors overvalue firms with bloated balance sheets?" *Journal of Accounting and Economics*, 38, 297–331.

Huberman, G. and Regev, T., 2001, "Contagious speculation and a cure for cancer: a nonevent that made stock prices soar," *Journal of Finance*, 56, 387–396.

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

Jung, J. and R.J. Shiller, 2005, "Samuelson's dictum and the stock market," *Economic Inquiry*, 43 (2), 221–228.

Jurafsky, D. and J. H. Martin, 2008, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Second Edition, Prentice Hall.

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.

Manela, A., and A. Moreira, 2015, "News implied volatility and disaster concerns." Available at ssrn.com/abstract=2382197.

Manning, C.D. and Schütze, H., 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Maćkowiak, B, and M. Wiederholt, 2009a, "Optimal sticky prices under rational inattention," Working Paper 1009, European Central Bank.

Maćkowiak, B, and M. Wiederholt, 2009b, "Optimal sticky prices under rational inattention," *American Economic Review*, 99, 769–803.

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

Pfaff, B., 2008, "VAR, SVAR and SVEC models: implementation within R package vars," *Journal of Statistical Software*, 27, 1–32.

Sicherman, N., G. Lowenstein, D.J. Seppi, and S. Utkus, 2015, "Rational attention," *Review of Financial Studies*, to appear.

Sims, C., 2003, "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Sims, C., 2015, "Rational inattention and monetary economics," *Handbook of Monetary Policy*, Elsevier, in press.

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

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. 2011, "All the news that's fit to reprint: Do investors react to stale information?" *Review of Financial Studies*, 24(5), 1481–1512.

1	Berkshire Hathaway	26	Australia & New Zealand Bank
2	Wells Fargo	27	AIG
3	Ind & Comm Bank of China	28	BNP Paribas
4	JP Morgan Chase	29	National Australia Bank
5	China Construction Bank	30	Morgan Stanley
6	Bank of China	31	Itau Unibanco
7	HSBC Holdings	32	UBS
8	Agricultural Bank of China	33	Bank of Communications
9	Bank of America	34	Royal Bank of Scotland
10	Visa	35	Prudential
11	China Life Insurance	36	Simon Property Group
12	Citigroup	37	Barclays
13	Commonwealth Bank of Australia	38	Bank of Nova Scotia
14	Ping An Insurance	39	Blackrock
15	Mastercard	40	AXA
16	Banco Santander	41	Banco Bilbao Vizcaya Argentaria
17	Westpac Bank	42	China Merchants Bank
18	American Express	43	Metlife
19	Royal Bank of Canada	44	Banco Bradesco
20	Lloyds	45	Nordea Bank
21	Goldman Sachs	46	Zurich Insurance
22	Mitsubishi UFJ	47	Intesa Sanpaolo
23	US Bancorp	48	ING
24	Allianz	49	Sumitomo Mitsui
25	TD Bank	50	Allied Irish Banks

Table 1: Companies included in the Thomson Reuters news sample.

	Avg mkt cap (usd)	Percent of all articles	Number of firms
UNITED STATES	137.47	44.25	15
BRITAIN	82.73	19.11	5
AUSTRALIA	70.45	6.35	4
CANADA	72.45	6.08	3
SPAIN	68.28	4.68	2
FRANCE	70.59	4.63	2
NETHERLANDS	55.70	3.19	1
CHINA	136.00	2.70	8
GERMANY	80.20	2.22	1
SWITZERLAND	57.28	1.95	2
JAPAN	72.26	1.69	2
IRELAND	41.84	1.04	1
ITALY	57.52	0.80	1
BRAZIL	37.46	0.68	2
SWEDEN	45.87	0.63	1

Table 2: Companies are grouped by country of domicile. Within each country, the table shows the average market capitalization of the companies in the sample as of November 2015. Also shown are the percent of all articles in the Thomson Reuters dataset that mention companies from a particular country of domicile, as well the number of firms classified as being domiciled in a given country.

	ENTNEG	ENTPOS	ENTALL	SENTNEG	SENTPOS
Mean correlation	0.197	-0.003	0.095	0.309	-0.102
S.E.	0.026	0.019	0.024	0.024	0.017

Table 3: Correlation and standard error between different entropy and sentiment measures and 1 month at-the-money implied volatilities for the 50 stocks in our sample, and for the aggregate level sentiment and entropy series. The aggregate entropy series used here are the ones derived from the list of n-grams from all articles in month t, and not from the first principal component of the single name series. So each correlation is an average across 51 observations. If a stock implied volatility series is not present, and for the aggregate measures, the VIX index is used instead of single name implied volatility. Cross-sectional standard errors, which assume independence, are shown.

Month	Year	w1	w2	w3	w4	Total	Rank	p_i	m_i
9	2008	nyse	order	imbalance	_mn_	81	1	0.009	0.020
9	2008	the	collapse	of	lehman	38	2	0.004	0.004
9	2008	filed	for	bankruptcy	protection	138	3	0.016	0.245
9	2008	problem	accessing	the	internet	33	400	0.004	0.961
9	2008	imbalance	_n_	shares	on	299	401	0.034	0.999
9	2008	order	imbalance	_n_	shares	299	402	0.034	0.999
5	2012	_bn_	from	a	failed	28	1	0.008	0.009
5	2012	the	euro	zone	crisis	36	2	0.011	0.087
5	2012	declined	to	comment	on	56	3	0.017	0.258
5	2012	you	experience	problem	accessing	77	208	0.023	0.998
5	2012	experience	problem	accessing	the	77	209	0.023	0.998
5	2012	problem	accessing	the	internet	77	210	0.023	0.998

Table 4: This table shows the top and bottom three 4-grams, as determined by their contribution to $ENTNEG$ in selected months of our sample. The "Total" column shows the number of times the given n-gram has appeated in that month, and the "Rank" column gives its rank by entropy contribution – this is lower than 5000 because we restrict analysis to those n-grams which are classified as having negative sentiment. p_i and m_i are the in-sample probability and the training sample conditional probability for the n-gram (see equation (6)). Note that some of the 4-grams come from the same 5-gram.

	ENTNEG	ENTPOS	ENTALL	SENTNEG	SENTPOS	ENTSENT_NEG	VIX	SPX_rvol
Mean	7.401	6.443	7.446	3.744	2.233	28.156	21.169	17.366
Min	2.140	2.172	4.724	1.484	0.819	7.881	10.420	6.310
Max	11.592	16.747	9.908	8.077	4.958	77.348	59.890	79.190
SD	1.837	2.092	1.066	1.265	0.594	13.948	7.876	9.892

Table 5: This table reports summary statistics for the aggregate news-based measures, as well as the VIX and realize volatility for S&P 500. Start and End refer to the start and end dates of data availability for the variable in question. $SENTNEG$ and $SENTPOS$ are aggregate negative and positive sentiment measures. $ENTALL$, $ENTNEG$ and $ENTPOS$ are the first principal components of single-name level entropy measures applied to all n-grams, and those classified as negative and positive respenctively. $ENTSENT\_NEG$ interacts $SENTNEG$ with $ENTNEG$. All data series are monthly, and run from April 1998 to December 2014.

	SENTNEG	SENTPOS	ENTALL	ENTNEG	ENTPOS	ENTSENT_NEG	VIX
SENTNEG	1.00						
SENTPOS	-0.14	1.00					
ENTALL	-0.18	-0.42	1.00				
ENTNEG	0.19	-0.44	0.71	1.00			
ENTPOS	-0.09	-0.16	0.56	0.34	1.00		
ENTSENT_NEG	0.86	-0.32	0.19	0.64	0.08	1.00	
VIX	0.46	-0.37	0.30	0.48	0.15	0.60	1.00

Table 6: This table reports contemporaneous correlations among monthly levels of our news-based indicators and the VIX index. $SENTNEG$ and $SENTPOS$ are aggregate negative and positive sentiment measures. $ENTALL$, $ENTNEG$ and $ENTPOS$ are the first principal components of single-name level entropy measures applied to all n-grams, and those classified as negative and positive respenctively. $ENTSENT\_NEG$ interacts $SENTNEG$ with $ENTNEG$.

	1	2	3	4
Num Lags	6	3	3	3
Lag IVOL	false	false	true	true
Fwd Step	0	0	0	1
ENTNEG	0.09***	0.063***	-0.001	0.003
ENTPOS	-0.052**	-0.047**	-0.016**	-0.008
ENTALL	-0.028	-0.031**	0	0.001
SENTNEG	0.138***	0.116***	0.005**	0.009***
SENTPOS	-0.019	-0.024*	0.001	0.006
ENTSENT_NEG	0.221***	0.177***	0.004	0.013***
ENTSENT_POS	-0.052**	-0.035**	-0.011	0
3 var	0.181***	0.145***	0.005	0.009
1 var Control mean R2	0.105	0.0796	0.69	0.549
3 var Control mean R2	0.28	0.18	0.711	0.595

Table 7: This table reports the results of estimating equation (9) in the text. "Num Lags" refers to the order of the lag operator $\mathcal{L}_s$, "Lag IVOL" refers to whether lagged values of $IVOL^j_{1mo}$ are included on the right hand side, and "Fwd Step" is the value of ϕ. We regress single-name implied volatility in month t on lags of each of the news-based measures in this table (the evaluation regression). We then repeat the same regression using lags of $ARTICLE\_PERCTOT$ as the regressor (the control regression). We run these regressions for each single-name in our sample, and collect the R^2's across all single-name regressions. The "3 var" model uses $SENTNEG$, $ENTNEG$, and $ENTSENT\_NEG$ as regressors (as described in Section 4.1). We then measure the area under the R^2 curve ($\Pr(R^2 > x)$) for the control and evaluation regressions. This table reports the differences in the areas under the two curves (equal to the average of the pairwise R^2 differences). Standard errors are obtained by assuming the pairwise differences in R^2's between the control and evaluation regressions are independent across all names. '*', '**', and '***' indicate significance at the 0.10, 0.05, and 0.01 levels respectively. The "Control mean R2" rows show the mean of the cross-sectional R^2's from the control regressions.

	SENTPOS	ENTPOS	ENTSENT_POS	SENTNEG	ENTNEG	ENTSENT_NEG	ARTICLE_PERCTOT
ivol_l1	-0.001	0.001	0.000	0.001	0.001	0.001	0.001
ivol_l2	-0.001	-0.001	-0.001	0.000	-0.003***	-0.002*	0.003***
rvol_l1	0.000	0.001	0.001	0.004***	0.004***	0.006***	0.000
rvol_l2	0.000	0.000	0.000	-0.001	0.001	0.000	-0.002***
ret_mi_l1	-0.007**	-0.007**	-0.007**	0.003	0.009***	0.007**	0.007***
ret_mi_l2	0.000	-0.005	-0.005	0.003	0.006**	0.007**	-0.001
SENTPOS_l1	0.146***						
SENTPOS_l2	0.095***						
ENTPOS_l1		0.150***					
ENTPOS_l2		0.140***					
ENTSENT_POS_l1			0.131***				
ENTSENT_POS_l2			0.075***				
SENTNEG_l1				0.221***			
SENTNEG_l2				0.169***			
ENTNEG_l1					0.195***		
ENTNEG_l2					0.151***		
ENTSENT_NEG_l1						0.195***	
ENTSENT_NEG_l2						0.113***	
ARTICLE_PERCTOT_l1							0.366***
ARTICLE_PERCTOT_l2							0.217***
Sum Last Two	0.241***	0.29***	0.207***	0.39***	0.346***	0.307***	0.583***
R2 adj	0.042	0.052	0.028	0.154	0.136	0.166	0.264

Table 8: This table reports the results of the panel model from (12). The dependent variable is shown in the column heading, with the regressors in the rows. The row labeled "Sum Last Two" shows the sum of the two bottom-most coefficients in each column. The regression is run with individual fixed effects. Residuals are clustered by time for computing standard errors. '*', '**', and '***' indicate significance at the 0.10, 0.05, and 0.01 levels respectively.

	ivol (1)	ivol (2)	ivol (3)	ivol (4)	ivol (5)	ivol (6)	ivol (7)
ivol_l1	0.307***	0.305***	0.305***	0.304***	0.252***	0.251***	0.289***
ivol_l2	0.094***	0.080	0.080	0.091***	0.103***	0.102***	0.067***
rvol_l1	0.209***	0.209***	0.208***	0.207***	0.226***	0.223***	0.223***
rvol_l2	0.102***	0.108***	0.107***	0.098***	0.104***	0.100***	0.119***
ret_mi_l1	0.364***	0.375***	0.374***	0.359***	0.420***	0.411***	0.371***
ret_mi_l2	-0.175***	-0.166***	-0.167***	-0.173***	-0.173***	-0.174***	-0.165***
ARTICLE_PERCTOT_l1	-0.678**	-0.666*	-0.638*	-0.732**	-0.806**	-0.913***	-0.681**
ARTICLE_PERCTOT_l2	0.270	0.148	0.168	0.233	0.149	0.247	0.224
SENTPOS_l1	-0.422**						
SENTPOS_l2	-0.235						
ENTPOS_l1		-0.369					
ENTPOS_l2		-0.395					
ENTSENT_POS_l1			-0.635***				
ENTSENT_POS_l2			-0.406				
SENTNEG_l1				0.802***			
SENTNEG_l2				0.467*			
ENTNEG_l1					0.225		
ENTNEG_l2					0.757***		
ENTSENT_NEG_l1						0.739***	
ENTSENT_NEG_l2						0.807***	
Sum Last Two	-0.657**	-0.764**	-1.041***	1.269***	0.982***	1.546***	-0.456
R2 adj	0.576	0.576	0.576	0.577	0.593	0.594	0.574

Table 9: This table reports the results of the panel model from (11). The dependent variable is shown in the column heading, with the regressors in the rows. The row labeled "Sum Last Two" shows the sum of the two bottom-most coefficients in each column. The regression is run with individual fixed effects. Residuals are clustered by time for computing standard errors. '*', '**', and '***' indicate significance at the 0.10, 0.05, and 0.01 levels respectively.

	rvol (1)	rvol (2)	rvol (3)	rvol (4)	rvol (5)	rvol (6)	rvol (7)
ivol_l1	0.269***	0.265***	0.264***	0.264***	0.295***	0.292***	0.245***
ivol_l2	0.120***	0.128***	0.128***	0.117***	0.118***	0.118***	0.091***
rvol_l1	0.332***	0.312***	0.311***	0.330***	0.311***	0.306***	0.349***
rvol_l2	0.122***	0.130***	0.130***	0.116***	0.116***	0.108***	0.140***
ret_mi_l1	0.859***	0.928***	0.925***	0.850***	0.860***	0.845***	0.864***
ret_mi_l2	-0.093	-0.109	-0.114	-0.090	-0.106	-0.105	-0.078
ARTICLE_PERCTOT_l1	-1.247***	-1.050**	-0.975**	-1.315***	-1.322***	-1.433***	-1.270***
ARTICLE_PERCTOT_l2	0.026	-0.244	-0.254	-0.048	-0.147	-0.031	-0.024
SENTPOS_l1	-0.684**						
SENTPOS_l2	-0.400						
ENTPOS_l1		0.051					
ENTPOS_l2		-0.431					
ENTSENT_POS_l1			-0.556*				
ENTSENT_POS_l2			-0.662*				
SENTNEG_l1				1.646***			
SENTNEG_l2				0.233			
ENTNEG_l1					0.986***		
ENTNEG_l2					0.521*		
ENTSENT_NEG_l1						1.977***	
ENTSENT_NEG_l2						0.451	
Sum Last Two	-1.085***	-0.379	-1.217***	1.878***	1.507***	2.428***	-1.294***
R2 adj	0.589	0.583	0.584	0.591	0.592	0.595	0.587

Table 10: This table reports the results of the panel model from (11). The dependent variable is shown in the column heading, with the regressors in the rows. The row labeled "Sum Last Two" shows the sum of the two bottom-most coefficients in each column. The regression is run with individual fixed effects. Residuals are clustered by time for computing standard errors. '*', '**', and '***' indicate significance at the 0.10, 0.05, and 0.01 levels respectively.

Model	Alpha	Mkt_RF	SMB	HML	WML	Adj R2
NEG[N=6,f=0.04]	0.454	0.066	0.230	0.126	0.040	-0.009
	(0.95)	(0.57)	(1.24)	(0.69)	(0.36)	
NEG[N=6,f=0.04] w/ ENTALL	0.634	-0.019	0.185	0.189	0.046	-0.009
	(1.40)	(-0.19)	(0.90)	(1.09)	(0.38)	
POSNEG[N=1,f=0.2]	0.185	0.032	0.163	0.059	-0.031	-0.012
	(0.96)	(0.56)	(1.38)	(0.27)	(-0.40)	
POSNEG[N=1,f=0.2] w/ ENTALL	0.4732	0.0117	0.1323	0.0073	-0.0447	-0.012
	(2.06)	(0.14)	(1.27)	(0.06)	(-0.52)	

Table 11: Results of monthly regressions of news-based zero-investment portfolio returns on the Fama-French global three factor model (market, size (SMB), and value (HML)), and a global momentum factor (WML). The portfolio formation criteria are explained in Section 6. Alphas are in percent (e.g. 0.50 means 50 basis points). Robust t-statistics are shown in parentheses, and are obtained using Newey-West standard errors with automatic lag selection, as implemented in the `sandwich` package in **R**.

Fraction	0.04	0.1	0.2
	Months held 1		
SR	0.3785	0.2444	0.4173
SR_INT	0.4366	0.1102	0.3285
Diff SR	0.0581	-0.1341	-0.0887
Alpha	0.9487	0.3455	0.4720
Alpha_INT	1.0616	0.0987	0.3724
Diff Alpha	0.1129	-0.2467	-0.0996
	Months held 3		
SR	0.124	0.20214	0.2175
SR_INT	0.291	0.20099	0.1984
Diff SR	0.167	-0.00115	-0.0192
Alpha	0.221	0.11168	0.1072
Alpha_INT	0.474	0.12717	0.1322
Diff Alpha	0.253	0.01549	0.0249
	Months held 6		
SR	0.3747	0.3377	0.24552
SR_INT	0.4665	0.3739	0.23934
Diff SR	0.0919	0.0362	-0.00619
Alpha	0.4537	0.1907	0.10932
Alpha_INT	0.6342	0.3425	0.18966
Diff Alpha	0.1804	0.1518	0.08034

Table 12: For the **negative** sentiment sort, this table shows the annualized Sharpe ratios and monthly alphas from the four factor model described in Section 6. The table columns correspond to different sample fractions used in portfolio formation, and the rows correspond to different holding periods for the stocks in each portfolio. Each block shows the sentiment-only sort Sharpe ratios and alphas, the ones from the sentiment-entropy-interacted sort (labeled INT), as well as the difference of two (interacted minus non-interacted). Alphas are in percent (e.g. 0.50 means 50 basis points). Regressions are run with monthly data.

Fraction	0.04	0.1	0.2
	Months held 1		
SR	-0.142	0.0619	0.178
SR_INT	0.222	-0.0524	0.457
Diff SR	0.364	-0.1143	0.278
Alpha	-0.259	0.0345	0.185
Alpha_INT	0.418	-0.0582	0.473
Diff Alpha	0.677	-0.0927	0.288
	Months held 3		
SR	-0.2769	-0.246	-0.1016
SR_INT	-0.2537	-0.381	0.0665
Diff SR	0.0232	-0.136	0.1680
Alpha	-0.4975	-0.302	-0.1361
Alpha_INT	-0.5699	-0.453	0.0307
Diff Alpha	-0.0724	-0.151	0.1668
	Months held 6		
SR	-0.20735	-0.12945	-0.1393
SR_INT	-0.19877	-0.17426	0.0144
Diff SR	0.00857	-0.04481	0.1537
Alpha	-0.41354	-0.16644	-0.1349
Alpha_INT	-0.40896	-0.17037	0.0196
Diff Alpha	0.00459	-0.00393	0.1544

Table 13: For the **positive vs negative** sentiment sort, this table shows the annualized Sharpe ratios and monthly alphas from the four factor model described in Section 6. The table columns correspond to different sample fractions used in portfolio formation, and the rows correspond to different holding periods for the stocks in each portfolio. Each block shows the sentiment-only sort Sharpe ratios and alphas, the ones from the sentiment-entropy-interacted sort (labeled INT), as well as the difference of two (interacted minus non-interacted). Alphas are in percent (e.g. 0.50 means 50 basis points). Regressions are run with monthly data.

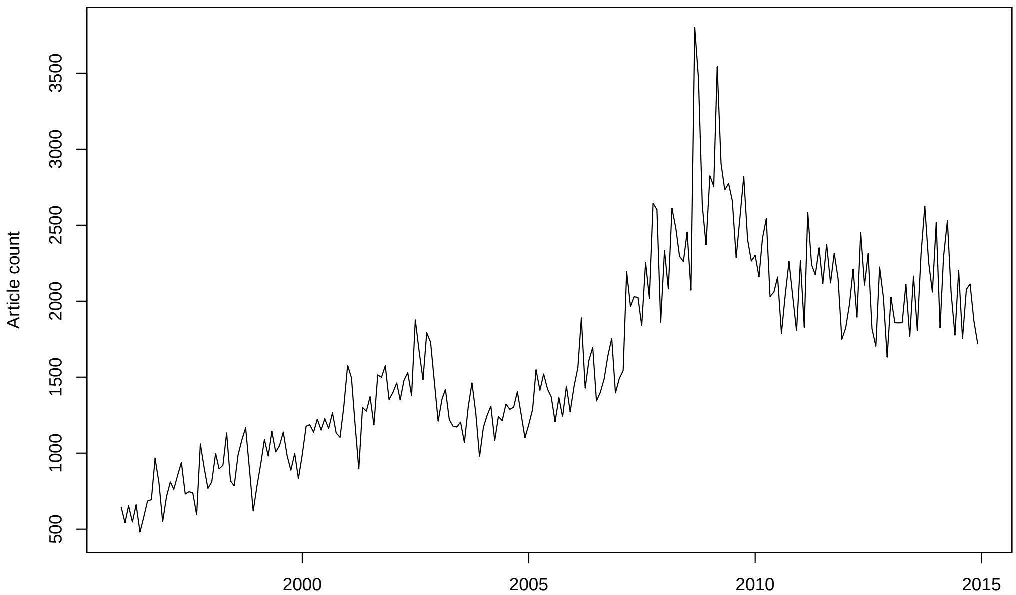

Figure 1: Monthly article count in the Thomson Reuters news sample.

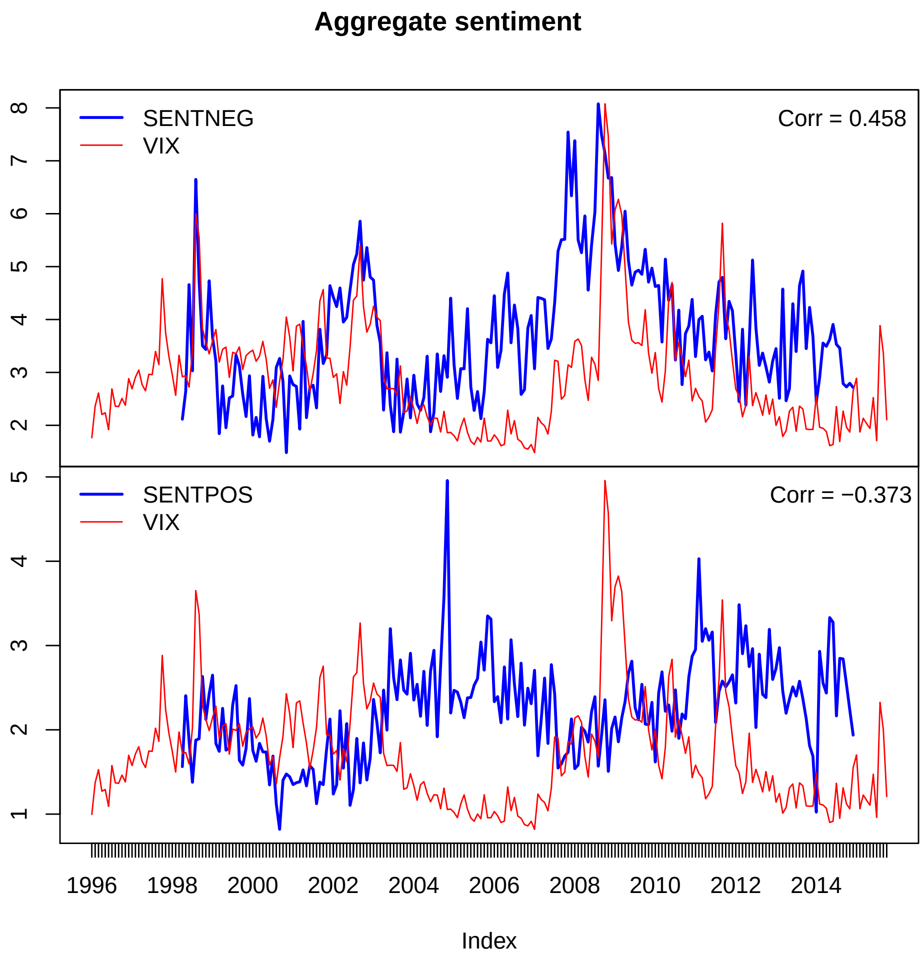

Figure 2: Monthly plots of $SENTNEG(t)$ and $SENTPOS(t)$ as defined in (7). Each series computes the proportion of all n-grams in a given month that are classified as having either positive or negative sentiment. Superimposed on each sentiment series is the scaled VIX index. Correlation between sentiment and VIX is shown in the upper right hand corner of each chart.

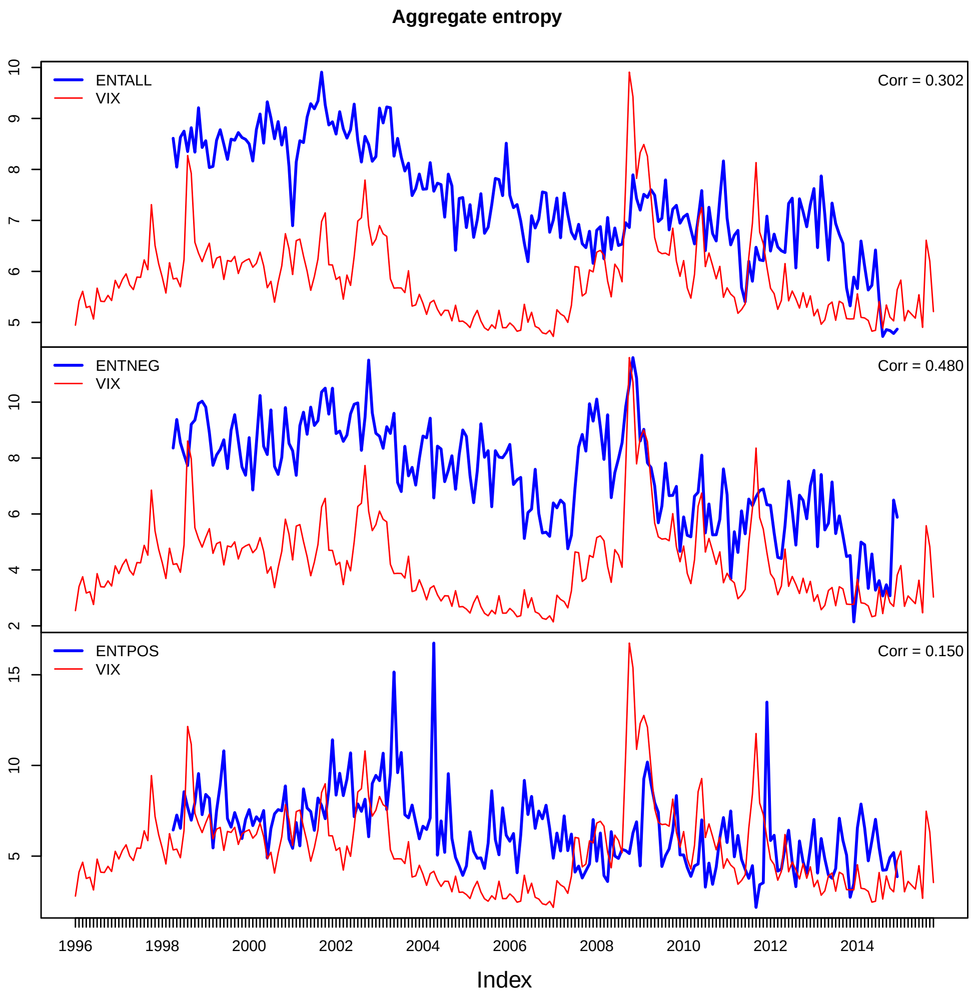

Figure 3: Monthly plots of $ENTALL(t)$, $ENTNEG(t)$ and $ENTPOS(t)$ as defined in Section 3.3. Each series is the first principal component of the associated single name entropy measures, for those names with observations available in all time periods of the sample. Superimposed on each entropy series is the scaled VIX index. Correlation between entropy and VIX is shown in the upper right hand corner of each chart.

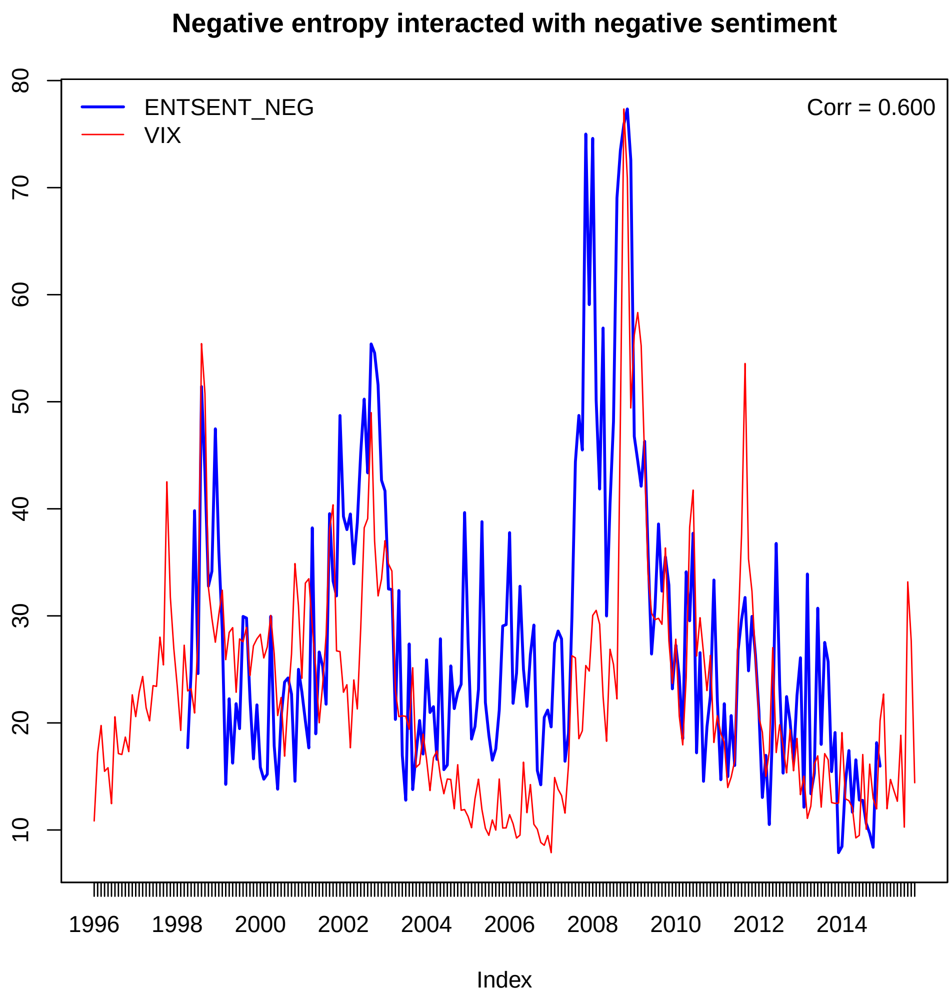

Figure 4: Monthly plot of $ENTSENT\_NEG(t) \equiv ENTNEG(t) \times SENTNEG(t)$. The entropy series is the first principal component of the associated single name entropy measures, for those names with observations available in all time periods of the sample. $SENTNEG$ is defined in (7). Superimposed on $ENTSENT\_NEG$ is the scaled VIX index. The correlation between $ENTSENT\_NEG$ and VIX is shown in the upper right hand corner.

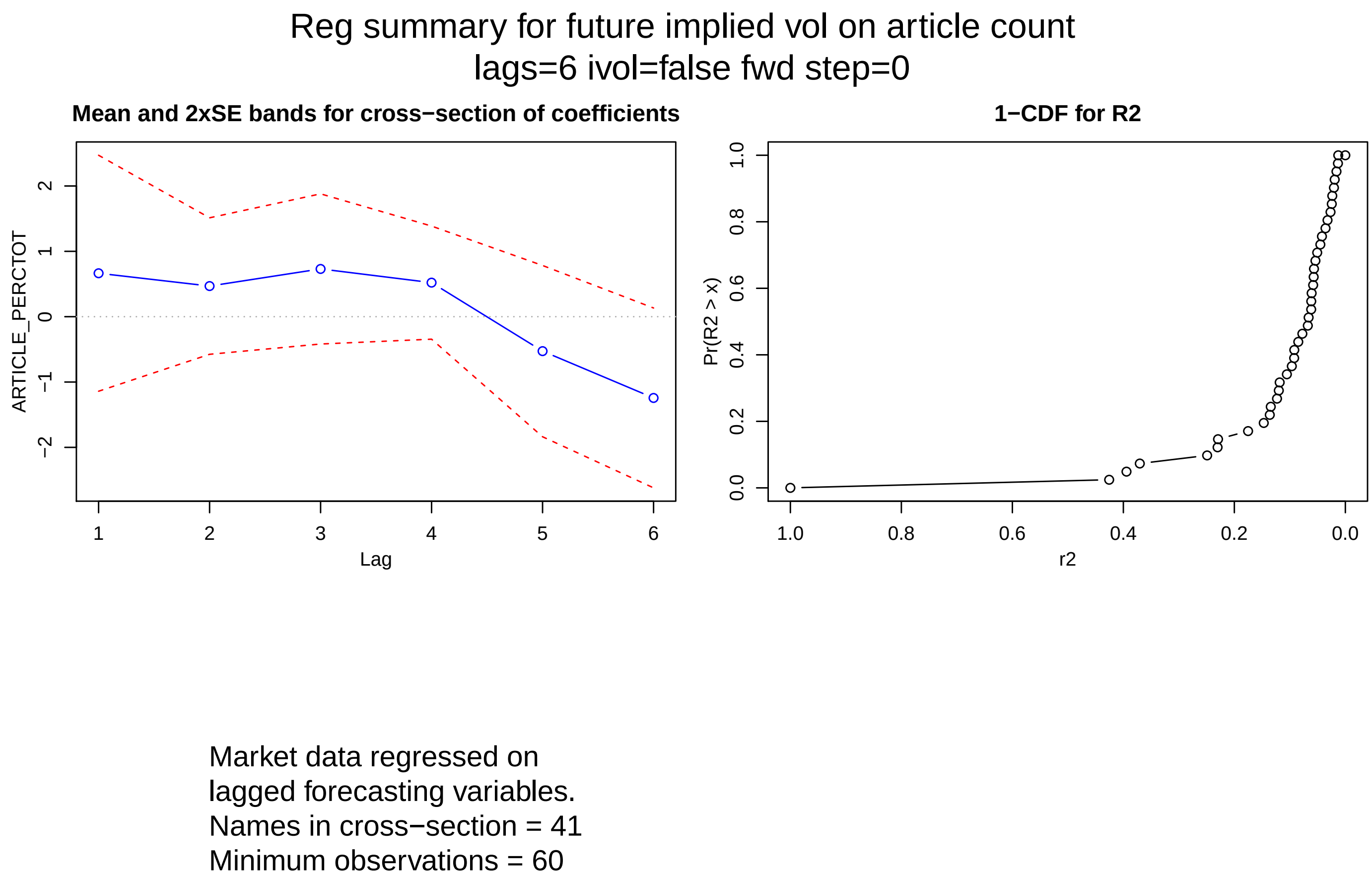

Figure 5: The results of the regression in (9) with $NEWS^j$ set to the percentage of articles in month t that mention company j. The $ARTICLE\_PERCTOT$ variables are normalized to have unit standard deviation. Shown are the cross-sectional mean of each coefficient with a two standard error band, and a plot of 1 minus the cumulative distribution function of the unadjusted R^2's from the single name regressions, i.e. $f(x) = \Pr(R^2 > x)$, with $ARTICLE\_PERCTOT$ as the right hand side variable. Note the x-axis starts at 1 and decreases to 0. Data are monthly.

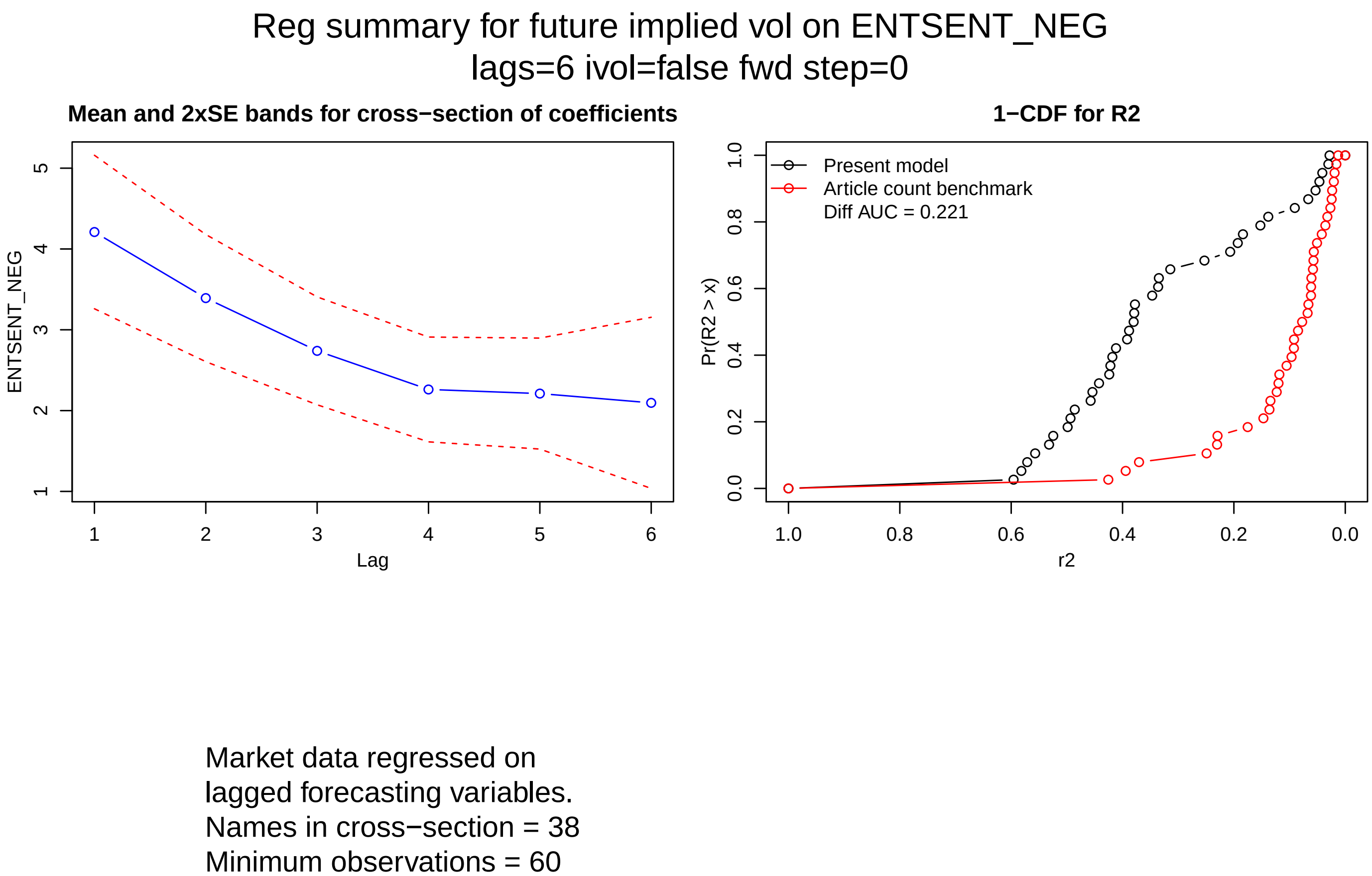

Figure 6: The results of the regression in (9) with $NEWS^j$ set to the month t interacted value of negative sentiment with negative entropy for company j. The $ENTSENT\_NEG$ variables are normalized to have unit standard deviation. Shown are the cross-sectional mean of each coefficient with a two standard error band, and a plot of 1 minus the cumulative distribution function of the unadjusted R^2's from the single name regressions, i.e. $f(x) = \Pr(R^2 > x)$, as well as the control R^2 curve for $ARTICLE\_PERCTOT$. Note the x-axis starts at 1 and decreases to 0. Data are monthly.

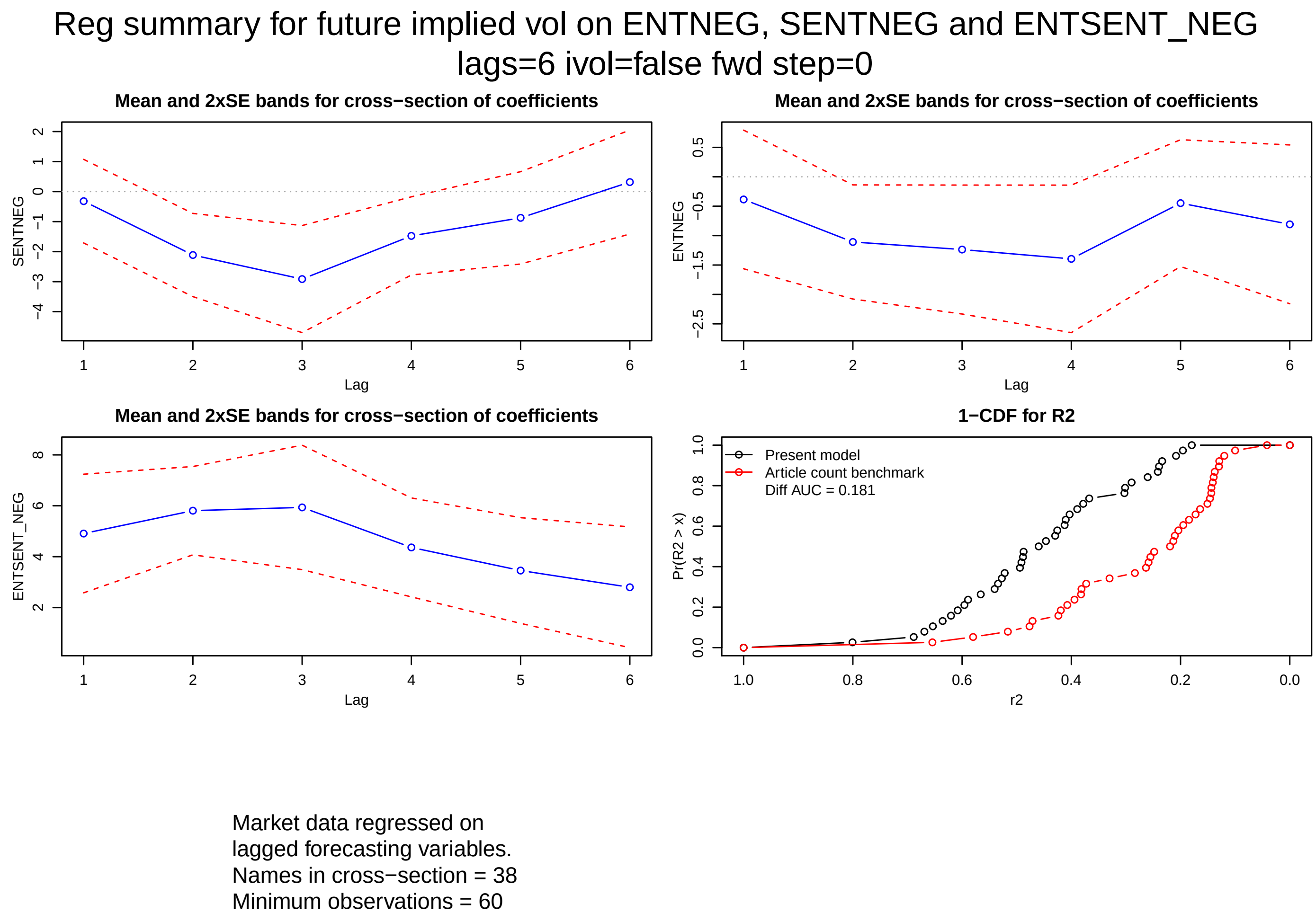

Figure 7: The results of the regression in (9) with three news-based variables: $SENTNEG$, $ENTNEG$ and $ENTSENT\_NEG$, all of which are normalized to have unit standard deviation. Shown are the cross-sectional mean of each coefficient with a two standard error band, and a plot of 1 minus the cumulative distribution function of the unadjusted R^2's from the single name regressions, i.e. $f(x) = \Pr(R^2 > x)$, as well as the control R^2 curve using $ARTICLE\_PERCTOT$, $NGRAM\_PERCTOT$ and the interaction term $ARTICLE\_PERCTOT \times NGRAM\_PERCTOT$ as the regressors. Note the x-axis starts at 1 and decreases to 0. Data are monthly.

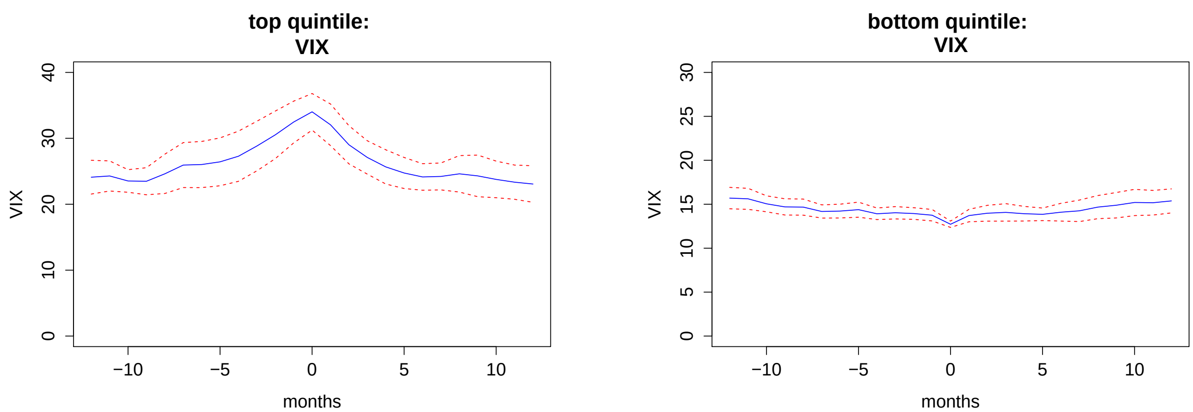

Figure 8: Average level of the VIX 12 months before and after high (left) and low (right) values of the VIX. High and low values are defined by the top and bottom quintiles. Dashed lines show plus and minus two standard errors.

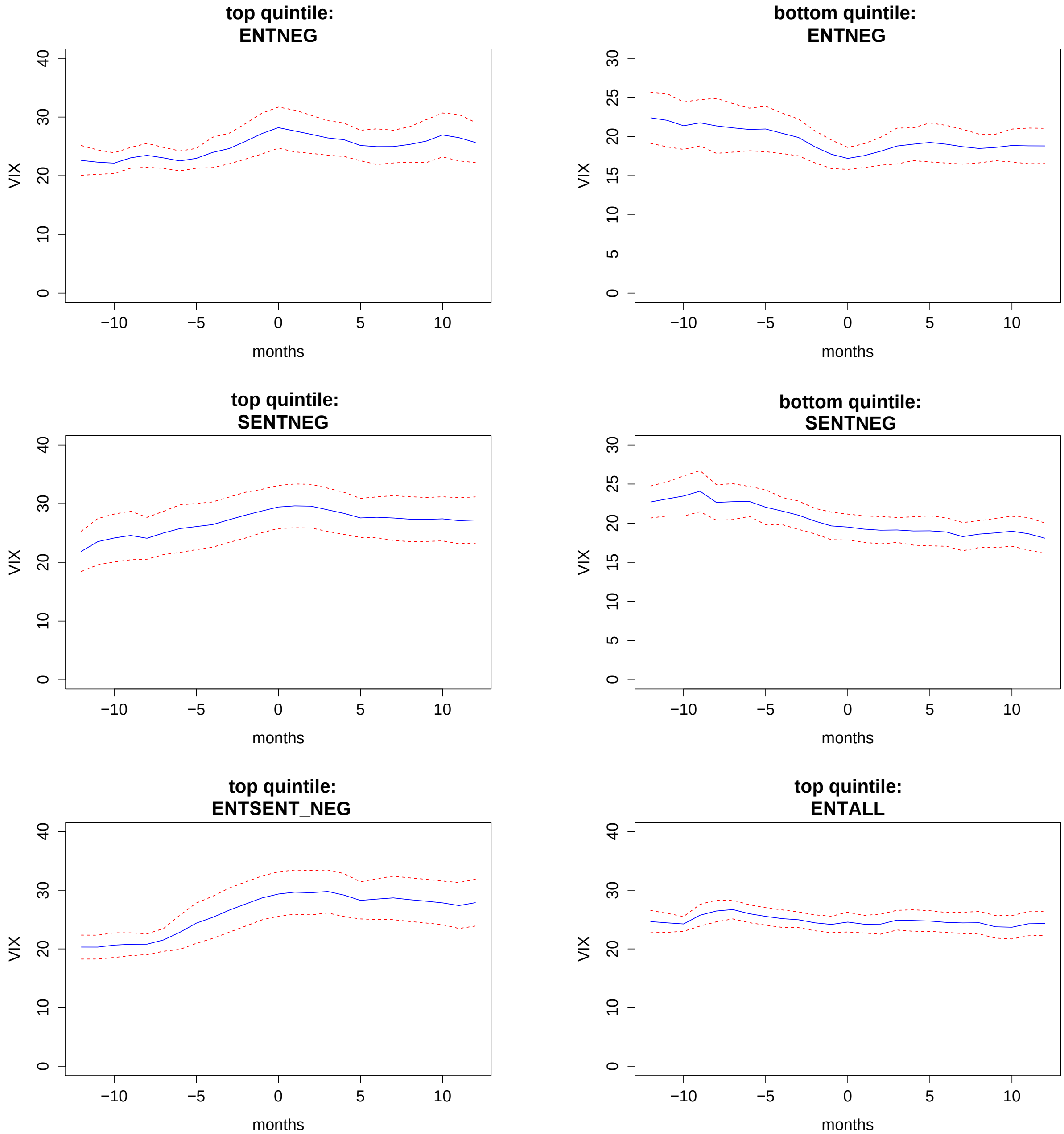

Figure 9: Average level of the VIX 12 months before and after high (left) and low (right) values of various entropy and sentiment measures. High and low values are defined by the top and bottom quintiles for each measure. Dashed lines show plus and minus two standard errors.

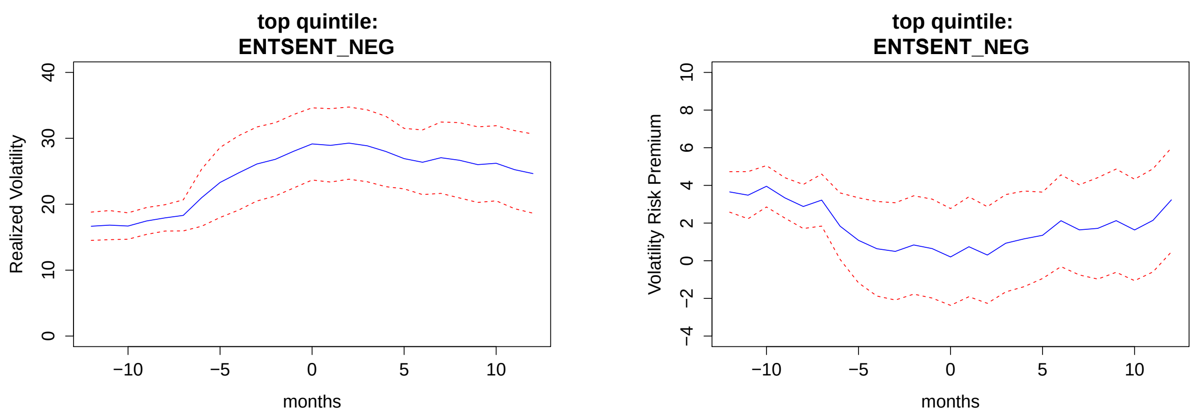

Figure 10: Average level of realized volatility (left) and the volatility risk premium (right) 12 months before and after top quintile values values of $ENTSENT\_NEG$. Dashed lines show plus and minus two standard errors.

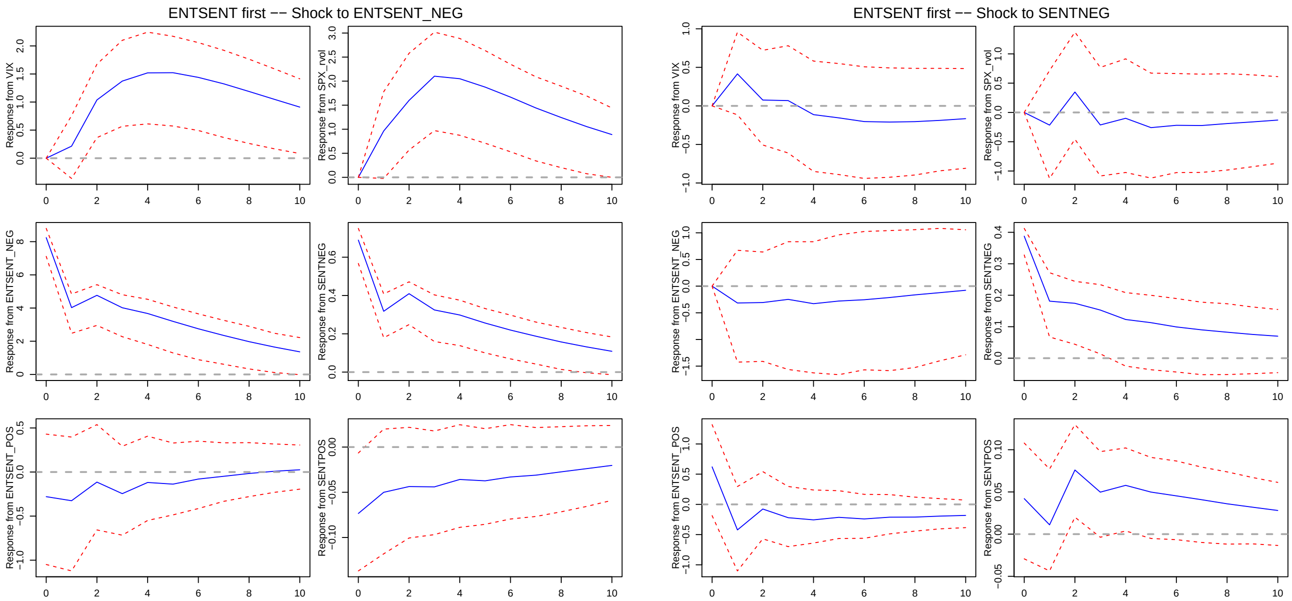

Figure 11: Impulse response functions for a shock to $ENTSENT\_NEG$ (left) and $SENTNEG$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

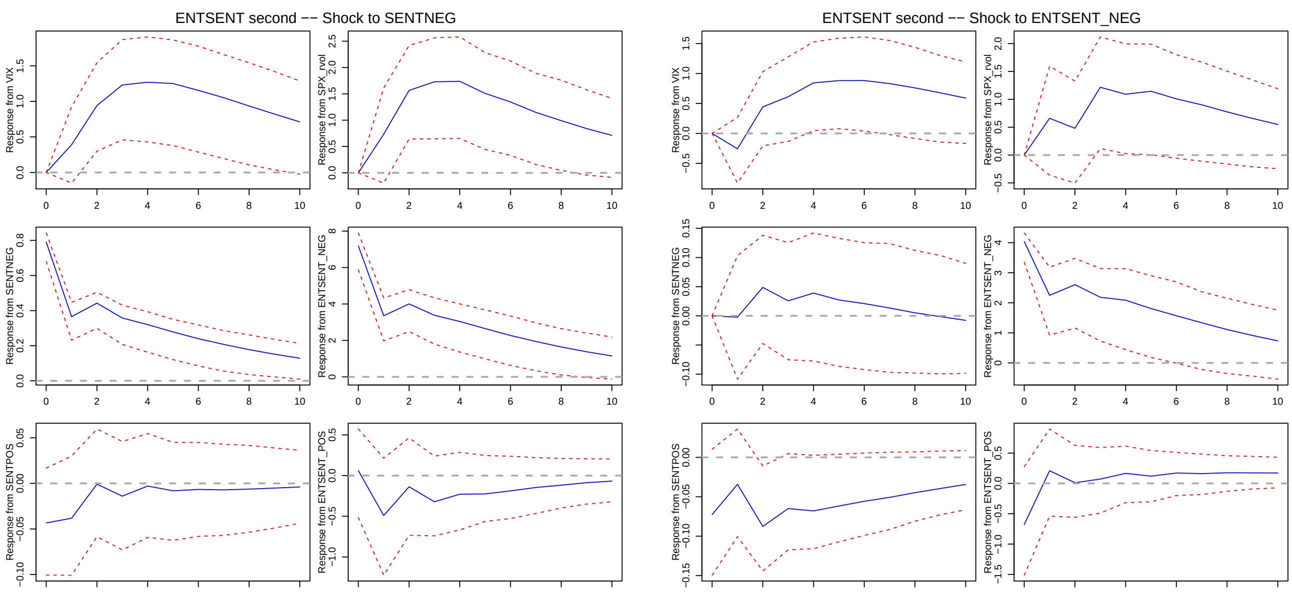

Figure 12: Impulse response functions for a shock to $SENTNEG$ (left) and $ENTSENT\_NEG$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

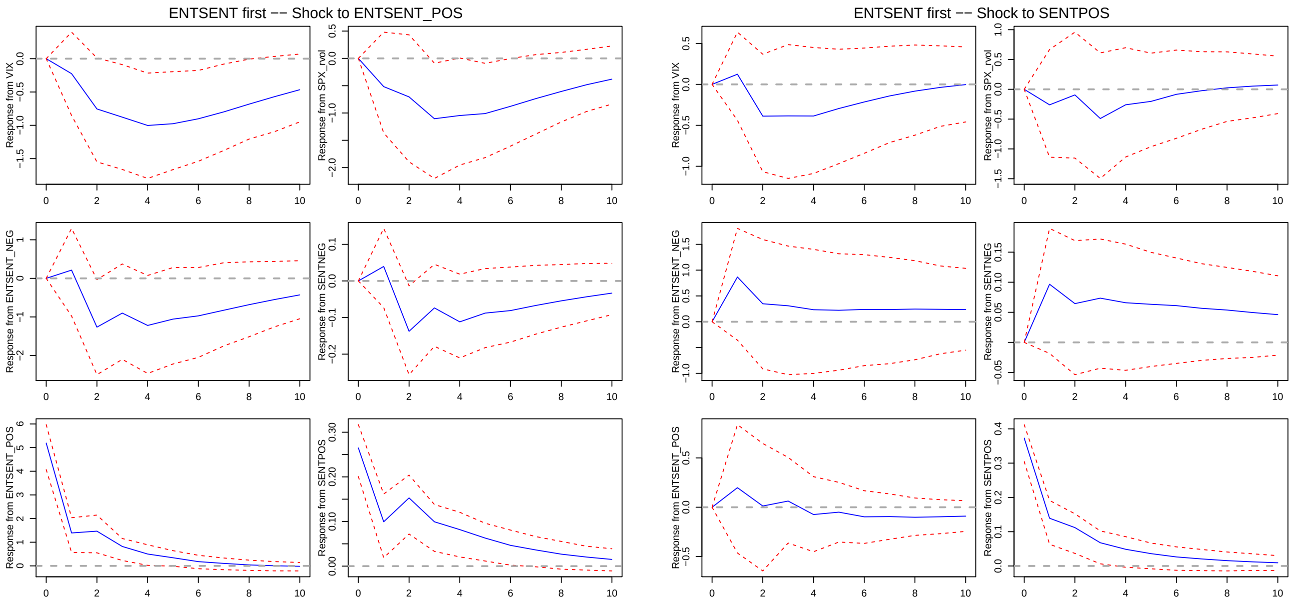

Figure 13: Impulse response functions for a shock to $ENTSENT\_POS$ (left) and $SENTPOS$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

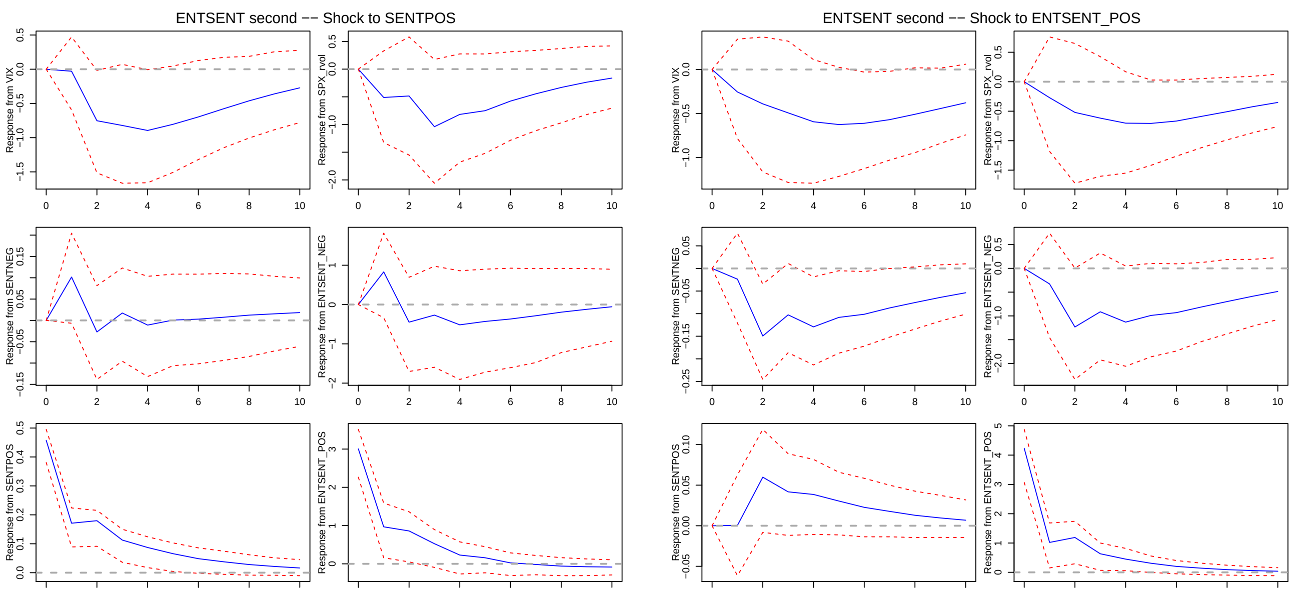

Figure 14: Impulse response functions for a shock to $SENTPOS$ (left) and $ENTSENT\_POS$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

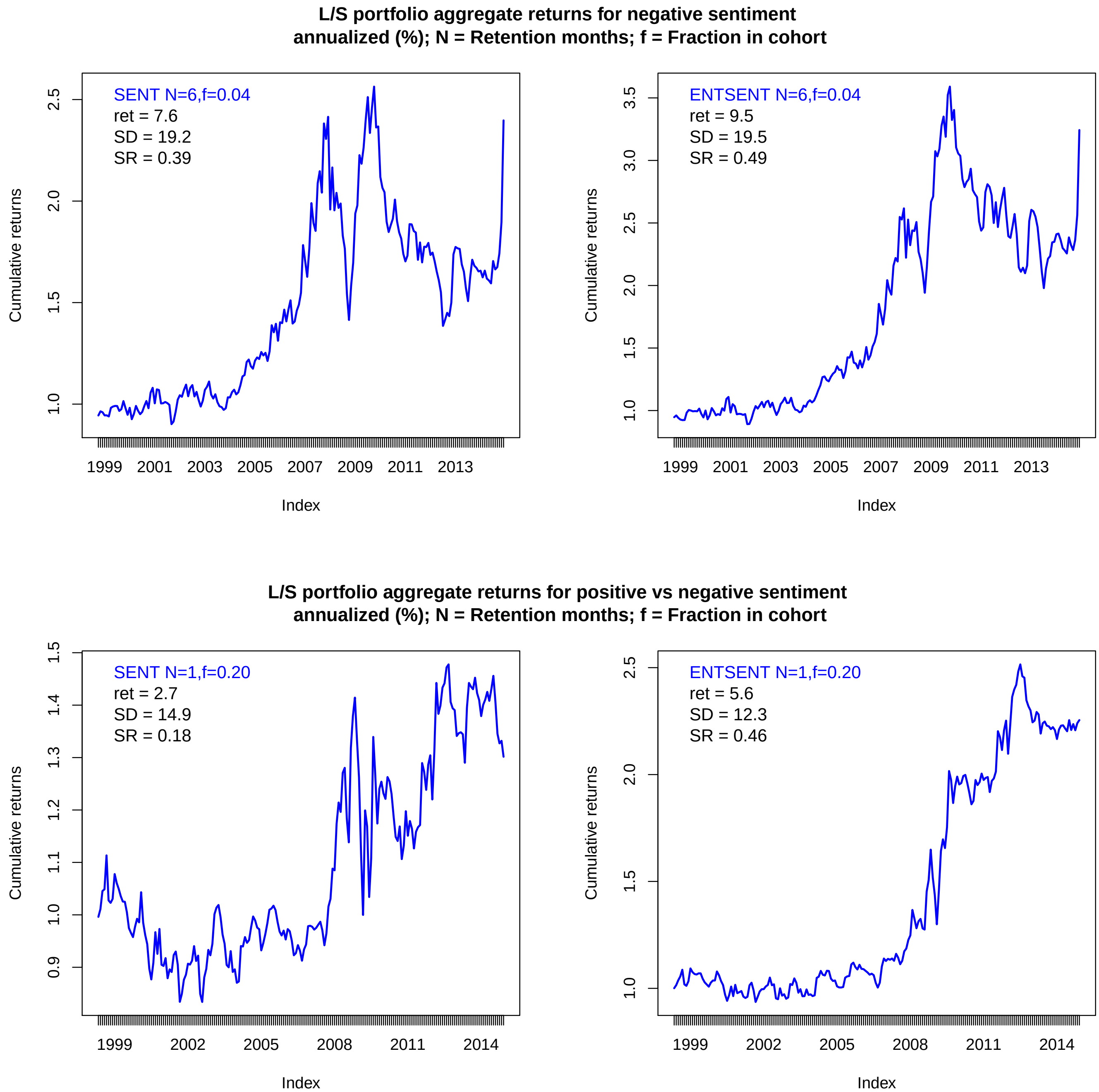

Figure 15: This figure shows the cumulative return of long-short portfolios formed using different news-based sorts. The long and short side of each portfolio contain the fraction f of all names for which returns are available in a given month. Data are monthly. Also shown are the arithmetic average monthly return on an annualized basis, as well as annualized return volatility (assuming uncorrelated monthly returns), and the annualized Sharpe ratio of each strategy. The sentiment sort $SENT$ and its entropy interacted version $ENTSENT$ are shown side by side. The four sorts are described in Section 6.

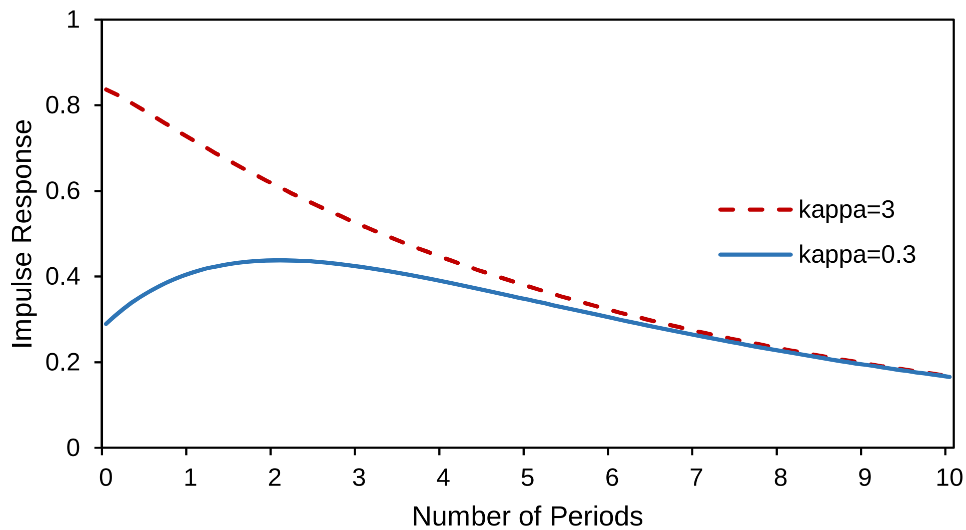

Figure 16: This figure shows impulse response functions in the model of Section 7. The response is hump-shaped for small κ (a tight information constraint) and monotonically decreasing for large κ. The other parameters are $\rho = 0.85$, $\lambda = 0$, and $a = 1$.