Does Unusual News Forecast Market Stress?

Paul Glasserman and Harry Mamaysky*

Initial version: July 2015. *Current version:* January 2017

Abstract

We find that an increase in the "unusualness" of news with negative sentiment predicts an increase in stock market volatility. Similarly, unusual positive news forecasts lower volatility. Our analysis is based on more than 360,000 articles on 50 large financial companies, published in 1996–2014. Unusualness interacted with sentiment forecasts volatility at both the company-specific and aggregate level. These effects persist for several months. Furthermore, unusual news is reflected in volatility more slowly at the aggregate than at the company-specific level. The observed behavior of volatility in our analysis can be explained by attention constraints on investors.

[*] Glasserman: Columbia Business School and Office of Financial Research (OFR), pg20@columbia.edu. Mamaysky: Columbia Business School, hm2646@columbia.edu. 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, Kent Daniel, Tara Sinclair, and seminar 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, the IAQF/Thalesians seminar, the Imperial College London Quantitative Finance Seminar, the Princeton Quant Trading Conference, the Columbia-Bloomberg Machine Learning in Finance Workshop, BNY Mellon's Machine Learning Day, the 2016 Philadelphia Fed Conference on Real-Time Data Analysis and the 2016 SIAM Financial Mathematics Conference 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. We also find that unusual news is reflected in volatility more slowly at the aggregate level than at the company-specific level, and we explore the causes and implications of this difference.

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), using company-specific news stories and responses, show this relationship also holds at the individual firm level. 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.[1] We document important differences between aggregate and company-specific responses to news – a novel finding that suggests greater efficiency at the micro level than the macro level. 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.

[1] Kogan et al. (2011) use the ability of text-based measures derived from companies' annual reports to forecast year-ahead volatility as a measure of the effectiveness of the Sarbanes-Oxley act.

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. News flow about these firms is particularly interesting to study because it contains important information about the macroeconomy. 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 sentiment yields the best predictor of future stock market volatility among the news measures we study.[2] Importantly, our analysis shows that news is not absorbed by the market instantaneously. We also find that the information in news articles relevant for future company-specific volatility is better reflected in firm-level option prices and realized volatility than macro-relevant information is reflected in the prices of S&P 500 options and S&P 500 realized volatility.

In simple forecasting regressions of company-specific implied volatility on lagged company-specific news measures, our interacted measure of unusual negative news, $ENTSENT\_NEG$, provides a statistically and economically significant predictor of volatility.[3] To control for known predictors of volatility (as documented, for example, in Poon and Granger 2003 and Bekaert and Hoerova 2014), we include lagged values of implied and realized volatilities and negative returns as explanatory variables in panel regressions of company-specific volatility measures on company-specific news measures. Even with the inclusion of the controls, our interacted measures of sentiment (both positive and negative) and unusualness remain economically and statistically significant at lags of up to two months, with positive measures forecasting a decrease

[2] Loughran and McDonald (2011) similarly find that sentiment provides a stronger signal when they put greater weight on less frequently occurring words, but the empirical word weights in Jegadeesh and Wu (2013) are only weakly related to word frequency.

[3] Negative sentiment and unusualness are also significant separately, but much less so.

in volatility and negative measures forecasting an increase. These results indicate that the information in our news measures is not fully reflected in contemporaneous option prices.

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, $ENTSENT$) is hump-shaped, peaking at around four months, and remaining significant even after ten months. This pattern suggests that the information in these news variables is absorbed slowly. We find similar results (but forecasting a decrease in volatility) for the interaction between positive sentiment and unusualness, though the effects are stronger for negative news than positive news.

To compare these macro results with micro results, we estimate a panel VAR for the corresponding company-specific variables — implied and realized volatility, positive and negative sentiment measures, and news measures interacted with unusualness. Impulse response functions again show that a shock to unusual negative (positive) news produces a statistically significant increase (decrease) in both implied and realized volatility. However, the responses now peak more quickly. When compared with our aggregate impulse response functions, these results indicate greater market efficiency at the micro level rather than the macro level.

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.[4] In contrast, we find that, even after controlling for known predictors of future volatility, news measures forecast volatility at lags as long as ten months at the macro level, and several months at the firm level. Furthermore, we show that company specific news is more fully incorporated into firm level option prices than aggregate news is incorporated into the price of S&P 500 option – even though both types of news are useful for forecasting future volatility. These are two of the most intriguing features of our results, so we suggest potential explanations.

The delayed response of volatility to news can be partly explained by a simple difference between forecasting volatility and forecasting returns. A predictable increase in realized volatil-

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

ity in the future — associated, for example, with a scheduled announcement or event — need not lead to a near-term increase in realized volatility. It should increase implied volatility, but arbitraging a predictable rise in volatility is 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. These arguments are fully consistent with market efficiency.

But this cannot be the entire story. Why does aggregate volatility initially underreact to news? Furthermore, why do individual firm options better reflect firm level news than S&P 500 options reflect aggregated news flow? One unifying explanation for these phenomena is that market participants have only a limited ability to process the information in tens of thousands of news articles. Investors who follow a specific company are better able to process news about that company than news about the macro economy because there is much less company-specific news and because the implications of company-specific news are often easier to grasp.

Attention allocation by investors offers a possible framework for these patterns. Several studies 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.

The capacity constraint faced by investors forces them to specialize – 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 by capacity constrained agents is examined in the models of Maćkowiak and Wiederholt (2009b), Peng and Xiong (2006), Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016), and Glasserman and Mamaysky (2016). Maćkowiak and Wiederholt (2009b) (with regard to firms) and Glasserman and Mamaysky (2016) (with regard to investors) find that economic agents favor micro over macro information.[5] The equilibrium effect of this, as shown in Glasserman and Mamaysky (2016), is to make prices more informationally efficient with regard to micro than macro information. Thus attention constrained investors are one explanation for why individual firm option prices better reflect company specific information than S&P 500 options reflect aggregate information.

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. Instead investors can only design a hedging instrument that tracks these unobservable macro state variables as closely as possible (in a sense made precise in Section 6). We interpret the VIX as an example of just such an imperfect hedging instrument. Next we solve for the price of this imperfect 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 monotonically decaying. This provides a coherent theoretical explanation for both the underreaction of aggregate volatility and the more timely reaction of single-name volatility that we find in the data.

Because the aggregate 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 (2016) develop an index of economic policy uncertainty based 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). See Das (2014)and Loughran and McDonald (2016) for an overview of text analysis with applications in finance.

The rest of this paper is organized as follows. Section 2 introduces the methodology we

[5] Peng and Xiong (2006) reach the opposite conclusion when looking at the information choice of a representative investor.

use, and Section 3 discusses the data and presents some summary statistics. Section 4 presents results based on company-specific volatility, and Section 5 examines aggregate volatility. Section 6 discusses possible explanations of our results and develops the connection with rational inattention. Section 7 concludes. An appendix presents evidence that the results in Sections 4 and 5 are robust across different subperiods of the data, and are not unduly influenced by the financial crisis of 2007-2009.

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.

An *n-gram* is a sequence of n words or, more precisely, n tokens.[6] Models that compute these types of probabilities are called n-gram models (in the above example, $n = 5$) because they give the probability of seeing the n^{th} word conditional on the first $n - 1$ words.

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

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

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.

Consider an N-word sentence $w_1 \ldots w_N$. We can write its probability as

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

N-gram models are used in this context to approximate conditional probabilities of the form $P(w_k|w_1 \ldots w_{k-1})$ 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 assume that

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

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

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

In (2), we have dropped first $n-1$ terms from (1).

A text or corpus is a collection of sentences.[7] 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 have to be estimated from a *training corpus*, which is usually very large relative to the evaluation text.

Assuming sentences are independent, the probability of an evaluation text is given by the product of the probabilities of its constituent sentences. Say the evaluation text consists of I distinct n-grams $\{w_1^i \ w_2^i \ w_3^i \ w_4^i\}$ each occuring c_i times. From (2), we see that the evaluation text probability can be written as

$$P_{eval} = \prod_{i=1}^{I} P(w_n^i|w_1^i \cdots w_{n-1}^i)^{c_i}. \tag{3}$$

[7] This has the effect of only counting as an n-gram a contiguous n-word phrase that does not cross a sentence boundary. Another option, as suggested by Jurafsky and Martin (2009, p.89), is to use a start and end of sentence token as part of the language vocabulary and then allow n-grams to lie in adjoining sentences. This would greatly increase the number of n-grams in our study, and due to data sparsity we did not pursue this option.

The probabilities $P(w_n^i|w_1^i \cdots w_{n-1}^i)$ in (3) are estimated from the training corpus. For a 4-gram $\{w_1\ w_2\ w_3\ w_4\}$, the empirical probability of w_4^i conditional on $w_1\ w_2\ w_3$ will be denoted by m_i, and is given by

$$m_i = \frac{\tilde{c}(\{w_1\ w_2\ w_3\ w_4\})}{\tilde{c}(\{w_1\ w_2\ w_3\})} \tag{4}$$

where $\tilde{c}(\cdot)$ is the count of the given 3- or 4-gram in the training corpus.[8]

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:

$$\begin{aligned} H_{eval} &\equiv -\frac{1}{\sum_k c_k} \log P(w_1 \dots w_N) = -\frac{1}{\sum_k c_k} \sum_{i=1}^{I} c_i \log m_i \\ &= -\sum_{i=1}^{I} p_i \log m_i, \end{aligned} \tag{5}$$

where p_i is the fraction of all n-grams represented by the i^{th} n-gram.

The evaluation text is *unusual* if it has low probability P_{eval}, relative to the training corpus. Equation (5) shows that, in an n-gram model, the evaluation text is unusual if there are n-grams 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_{eval} simply as the entropy of the evaluation text. Based on this definition, unusual texts will have high entropy.

The definition of entropy in (5) can apply to an arbitrary list of n-grams, as opposed to all the n-grams in a text. 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. The fraction of all n-grams represented by the i^{th} n-gram is therefore

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

[8] In Section A.1 of the Appendix we discuss how we handle the situation where a particular 4-gram was not observed in the training corpus.

Given a list of n-grams in month t, the entropy of that list will be defined as

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

where, as before, the m_i's are conditional probabilities estimated from a training corpus.

Alternative Measures

In their analysis of 10-Ks, Loughran and McDonald (2011) find that sentiment measures are more informative when individual words are weighted based on their frequency of occurrence. They use what is known in the text analysis literature as a *tf-idf* scheme because it accounts for term frequency and inverse document frequency. The weight assigned to each word in each document depends on the number of occurrences of the word in the document and the fraction of documents that contain the word. A word in a document gets greater weight if it occurs frequently in that document and rarely in other documents. This approach is less well suited to our setting because we do not analyze individual documents and because our unit of analysis is the n-gram rather than the individual word. The entropy measure allows a more direct measure of the unusualness of an entire body of text in one period relative to another.

Tetlock (2011) uses measures of similarity between news articles as proxies for staleness of news. His primary measure is the ratio of the number of words that two articles have in common to the number of distinct words occurring in the two articles combined. Although similar measures could potentially be used in our setting, Tetlock's (2011) approach seems better suited to comparing pairs of articles than to comparing large bodies of text.

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. To do so, researchers rely 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[9] word lists do a better job of sentiment categorization in a financial

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

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.

2.3 Variable definitions

Throughout the paper, our empirical work is at a monthly horizon, both for our news measures and our market and volatility data. In our analysis, we use a 4-gram model.[10]

Single-name entropy

We refer to the measure of unusualness of all month t articles about company j as $ENTALL^j(t)$. The entropy of the list of n-grams classified as being negative (positive) appearing in articles that mention company j is called $ENTNEG^j(t)$ ($ENTPOS^j(t)$). The p_i^j's from (6) come from these lists in month t, and the m_i's are estimated in a training corpus. 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.[11] More details about entropy calculations are in Section A.1 in the Appendix.

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

[11] 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 3.1) 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. When this phrase appears (11 times) in June 2012, we would still like to regard it as unusual.

Aggregate entropy

We find that the aggregate entropy measures calculated directly from a list of all month t n-grams can be unduly influenced by a small set of frequently occurring n-grams. 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 $m_i(t)$, 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, we define $ENTPOS$ 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.[12]

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 in month 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

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

with the analogous definition for $SENTNEG^j(t)$. Our aggregate measures of sentiment $SENTPOS$ and $SENTNEG$ are calculated using all month t n-grams.

3 Data

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.[13] Almost 90

[12] 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 $ENTALL$. 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.

[13] 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.

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.[14] 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 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.2 gives more details about our text processing procedure.

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 single name volatility series are 20-day realized volatilities of local currency returns, as calculated by Bloomberg. Our macro data series are the Chicago Board Options Exchange Volatility Index (VIX) and 20-day realized volatility for the S&P 500 Index computed by Bloomberg from daily returns.[15]

[14] All articles in a chain share the same *Article Id* code.

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

3.1 N-grams and their contribution to entropy

As an example, consider that 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 5,000 most frequently occurring 4-grams. In our 19 year dataset, we thus analyze $19 \times 12 \times 5000 = 1.14$mm 4-grams, of which 394,778 are distinct. The first three tokens in the latter represent 302,973 distinct 3-grams.

By sorting n-grams on their contribution, given by $-p_i \log m_i$, to the entropy of the overall month t corpus, we can identify for month t the most and least unusual 4-word phrases. Table 4 shows the three top and bottom phrases[16] 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

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

a meaningful, and potentially important, way.

3.2 Summary statistics

Table 3 shows the average contemporaneous correlation between the 50 individual volatility (realized and implied) and sentiment pairs. If an individual implied volatility series does not exist, we use the VIX as a stand-in. Cross-sectional standard errors are also calculated assuming independence of observations. We see that $SENTNEG^j$ ($SENTPOS^j$) is on average positively (negatively) correlated with single name volatility.

We observe a similar pattern at the aggregate level. 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 is 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 also shows the average correlation between the various single name entropy measures and single name implied or realized volatility. The average single name correlations for $ENTALL$ and $ENTNEG$ are positive, and the $ENTPOS$ average correlation is marginally negative though very close to zero.

Figure 3 shows the three aggregate entropy series ($ENTALL(t)$, $ENTNEG(t)$, and $ENTPOS(t)$), 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.

4 Single name volatility

To get a simple indication of whether our news-based measures are relevant to forecasting volatility, we can regress single name implied volatility (30-day at-the-money) on lagged values of each of the news measures. Doing so shows, for example, that the variable $ENTSENT\_NEG$ (defined as $ENTNEG \times SENTNEG$) is a statistically significant predictor at lags of up to six months. A one standard deviation increase in $ENTSENT\_NEG$ forecasts an increase in implied volatility of 3–4 percentage points, so the effect is economically as well as statistically significant. This interacted variable has a greater impact than either sentiment or unusualness alone. [17]

However, 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.

Poon and Granger (2003) suggest that the best performing volatility forecasting models include both implied and historical volatility as explanatory variables. They also point out that models using short-dated at-the-money implied volatility work about as well as more sophisticated approaches that take into account the volatility term structure and skew. Bekaert and Hoerova (2014) show that, at the index level, in addition to lags of implied and realized variance, stock price jumps also 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 portion of monthly returns r^- (labeled ret_mi in the tables) as explanatory variables for future realized and implied volatility.[18] Our basic specification for evaluating the forecasting power of a news-based measure $NEWS^j$ is the following panel regression:

$$\begin{aligned} VOL^j(t) &= a^j + c_1'\mathcal{L}_s RVOL^j_{30day}(t) + c_2'\mathcal{L}_s IVOL^j_{1mo}(t) + c_3'\mathcal{L}_s r^{-j}(t) \\ &\quad + b1'\mathcal{L}_s ARTPERC^j(t) + b2'\mathcal{L}_s NEWS^j(t) + \epsilon^j(t), \end{aligned} \tag{8}$$

where VOL is either either $IVOL$ or $RVOL$, a^j is an individual fixed effect term, $\mathcal{L}_s$ is an s-lag operator,[19]and $ARTPERC^j$ is the percent of all month t articles that mention company j. The variable $ARTPERC^j$ is intended to control for the information content of news volume. All news measures are normalized to have unit variance.

We show results for $s = 2$ (those for $s = 3$ are qualitatively similar and omitted to conserve space). We have run this specification in variance, log variance and volatility terms, and all of

[17] These results are available in the supplementary appendix.

[18] $r^- \equiv \max(-r, 0)$. Adding $r^+ \equiv \max(r, 0)$ as an explanatory variable was not impactful in any of our specifications, so we do not include this variable in our regression results.

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

these yield similar qualitative results. We show the volatility results in the paper because these are the easiest to interpret.

Before turning to the forecasting regression in (8), we examine briefly the drivers of our news-based measures. The following is our descriptive panel specification:

$$NEWS^j(t) = a^j + c_1'\mathcal{L}_2 RVOL^j_{30day}(t) + c_2'\mathcal{L}_2 IVOL^j_{1mo}(t) + c_3'\mathcal{L}_2 r^{-j}(t) + b'\mathcal{L}_2 NEWS^j(t) + \epsilon^j(t). \tag{9}$$

This is run with $NEWS^j$ set to each of the following:

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

Table 7 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 – though all the news measures exhibit some persistence.

Tables 8 and 9 show the results of the specification in (8) 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, though the magnitude of the effect is much smaller than that of $r^-(t-1)$).

Model 1 which has $ARTPERC$ 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. Along similar lines, Jiao, Veiga, and Walther (2016) find that idiosyncratic realized volatility is lower for companies that receive greater attention from the news media. Furthermore, this finding may shed light on Fang and Peress (2009) who show that stocks receiving high media attention earn lower returns in the ensuing month – perhaps part of the story is that such stocks are also less volatile.[20]

The positive category news measures (Models 2–4) all show up with negative coefficients

[20] Fang and Peress (2009) double sort by the prior month's media coverage and the prior month's idiosyncratic volatility, but do not study the effects of media coverage on future volatility.

(except in one case in Table 9), suggesting positive news at time $t-1$ or $t-2$ forecast lower time t implied and realized volatility, after controlling for known forecasting variables. Summing the lag 1 and lag 2 coefficients on $NEWS^j$, reported in the row labeled "Sum NM 2", allows us to evaluate the importance of the different 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 8 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 point drop (e.g. from 20 to 19) in implied volatility next month. The results for future realized volatility in Table 9 are similar.

The negative category news measures (Models 5–7) 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 (at both lags) in the latter implies a 1.546 (2.429) rise in next month's implied (realized) volatility (as can be seen from the "Sum NM 2" row of Tables 8 and 9) – again a very large economic effect.

In Model 8, we include all news based measures in the panel (except the non-interacted entropy measures). The results of this regression are very stark. When adding the two lags for both positive and negative sentiment ($SENTPOS$ and $SENTNEG$), the cumulative effect on future volatility is very close to zero, while the effects of $ENTSENT\_POS$ (on implied volatility) and $ENTSENT\_NEG$ (on both implied and realized volatility) are economically and statistically (when looking at the sum of coefficients of lag 1 and 2) very large[21] – in fact the sum of the two coefficients for both variables is comparable to the results from Models 2–7. Interestingly, unusual positive and unusual negative news *both* matter. A one standard deviation increase in unusual positive (negative) news over both lags leads to a drop (increase) in future realized and implied volatilities of between 1.3 – 2.9 points (e.g. from 15 to 17). This is a very large economic effect.

The bottom row of Tables 8 and 9 shows results from an F-test that restricts all text based variables (excluding $ARTPERC$) in each panel to be zero. For panels that forecast implied volatility (Table 8), we can reject the restriction for all models that include $SENTNEG$ or $ENTSENT\_NEG$, and marginally reject for models including only $ENSENT\_POS$ or $ENTNEG$. For Model 8, which includes all text measures except non-interacted entropy, the

[21] For Models 8 and 9, the row labeled "Sum NM 2" shows the sum and associated p-value for the coefficients on $ENTSENT\_POS$.

restriction is soundly rejected. For the realized volatility panels in Table 9, we can additionally reject the restriction for the $SENTPOS$ regression. The restriction that our text based measures are zero in Model 8 is again strongly rejected.

In summary, our panel results suggest that, even after controlling for known predictors of future volatility, our news based measures contain useful forecasting information. The coefficient estimates on lagged news-measures are 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 its own.

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 2.3, 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 interacted variables $ENTSENT\_NEG(t) = ENTNEG(t) \times SENTNEG(t)$ and $ENTSENT\_POS(t) = ENTPOS(t) \times SENTPOS(t)$. 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.

$SENTPOS$ and $ENTSENT\_POS$ have a negative correlation with the VIX, whereas all the other measures have a positive correlation. In particular, 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 constituents: 0.86 with $SENTNEG$

and 0.64 with $ENTNEG$. $ENTSENT\_POS$ is weakly negatively correlated with the VIX and with $ENTSENT\_NEG$.

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.

5.1 Impulse Response Functions

We investigate interactions among the aggregate variables through vector autoregressions (VARs). We estimate a VAR model in six variables, initially ordered as follows: VIX, SPX_RVOL (realized volatility), $SENTNEG$, $ENTSENT\_NEG$, $SENTPOS$, and $ENTSENT\_POS$.[22] 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 5 shows impulse response functions in response to a shock to $ENTSENT\_NEG$, together with bootstrapped 95 percent confidence intervals.[23] 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 response functions. The left panel of Figure 6 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

[22] 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.

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

$ENTSENT\_NEG$ following the order change. Taking Figures 5 and 6 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 7 and 8 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 7, left panel), whereas $SENTPOS$ does not (Figure 7, right panel). When the order of the variables is interchanged, $SENTPOS$ has a statistically significant effect on VIX (Figure 8, left panel), and $ENTSENT\_POS$ has a marginally significant effect on both VIX and realized volatility (Figure 8, 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 5. 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. As we will show in Section 6, single-name impulse responses estimated from a panel VAR operate over a shorter, but still multi-month time horizon. 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 5–8. 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.

As we argue in Section 6, such 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.

A robustness check

It is possible that the hump-shaped responses in our VARs are due to the fact that the news innovations in our sample are systematically about events that will take place four or five months in the future, and the VIX, since it only measures one-month ahead volatility, doesn't react to such news right away. To control for this possibility we rerun our VARs, but include the Mid-Term VIX (ticker VXMT) as the seventh variable (placed between the VIX and SPX_RVOL). The VXMT is constructed using S&P 500 options with 6-to-9 months left to expiration, and provides a six-month ahead volatility forecast. Though we have less data for this augmented VAR,[24] the shape of the impulse responses of the VIX, VXMT and SPX_RVOL to news innovations are qualitatively similar to our original specification. For example, the maximal response of the VXMT and of the VIX to an $ENTSENT\_NEG$ innovation occurs in month four (SPX_RVOL peaks in month three). This suggests that the market does not fully incorporate all relevant news about future volatility into the VXMT price. The results of this VAR are available in the supplementary appendix.

6 Interpreting the results

As we have shown in Sections 4 and 5, text based measures[25] are useful for forecasting future implied and realized volatility at the single name and macro levels, even after controlling for the information content of current and lagged implied and realized volatilities. This is despite the fact that there is a high contemporaneous correlation between text based measures and implied volatilities, at the single name and aggregate levels,[26] which suggests that a given month's implied volatilities, realized volatilities, and text based measures all respond to the same underlying uncertainty. One may argue that, even in an informationally efficient market, there

[24] Our data on the VXMT start in January 2008 (the VAR excluding VXMT runs from April 1998 to December 2015). From January 2008 to July 2016 the correlation between the VXMT and the 6 month 90% strike S&P 500 implied volatility (obtained from Bloomberg) is 99.52%. This S&P 500 implied volatility series starts in January 2005. We also estimate VARs where we replace the VXMT series with the S&P 500 implied volatility series in order to have more observations. The 90% strike VAR and the VXMT one – run over the shorter time window – yield almost identical results.

[25] In the ensuing discussion, we do not focus on the distinction between sentiment and unusualness.

[26] See Tables 3 (single name) and 6 (aggregate), as well as Figures 2, 3, and 4.

may be some information useful for forecasting future realized volatility that is not incorporated into current or lagged realized volatility. For example, such information may be explicitly about future – and not current – events. However, it is puzzling why implied volatility would not reflect such information.[27] In keeping with the interpretation by Tetlock et al. (2008) that their results were driven by "[i]nvestors [who] ... do not fully account for the importance of linguistic information about fundamentals," we conjecture that our results are driven by investors who do not fully account for the information content of language for future volatility.

6.1 Attention and News

Several studies have found evidence that the limits of human attention affect market prices. Dellavigna and Pollet (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 (2016) 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, and Solomon, Soltes, and Sosyura (2014) find that media coverage affects investors through salience rather than information. 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.

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.[28]

[27] Feinstein (1989) and Chernov (2007) showed that, for a class of stochastic volatility processes, at-the-money implied volatility of short-dated options is a good estimate of expected realized volatility over the life of the option as long as the volatility risk premium is either zero or constant. If the volatility risk premium is systematically related to our measures of news flow then this can partially account for some of our results. However, it is difficult to see how a risk premium story would be consistent with the full range of observations that we document in this section.

[28] 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

If investor inattention is indeed the mechanism underlying our results, we should find that when relevant information is harder to gather and interpret, and thus requires more investor attention, text based measures should contain more incremental forecasting ability. To investigate this further, we compare our single name results – where gathering and interpreting relevant information may be easier – from Section 4 to our aggregate results – where information is more voluminous and has implications that are harder to interpret – from Section 5. To make the analyses from these two sections comparable, we replicate our VAR specification for implied volatility (VIX) and realized volatility (SPX_RVOL) in our single name panels. The full regression model is given by

$$
\begin{aligned}
VOL^j(t) = \gamma_0^j &+ \gamma_1'\mathcal{L}_2 RVOL^j(t) + \gamma_2'\mathcal{L}_2 IVOL^j(t) + \\
&\gamma_3'\mathcal{L}_2 SENTPOS^j(t) + \gamma_4'\mathcal{L}_2 ENTSENT\_POS^j(t) + \\
&\gamma_5'\mathcal{L}_2 SENTNEG^j(t) + \gamma_6'\mathcal{L}_2 ENTSENT\_NEG^j(t) + \epsilon^j(t).
\end{aligned} \tag{10}
$$

The left hand side variable $VOL^j(t)$ is either realized or implied volatility for single names, and either VIX or SPX_RVOL for the aggregate series. Model 9 from Tables 8 and 9 shows estimates of (10) for the single names panels.[29] To complete the panel VAR model, we also run (10) with each of the four news variables on the left.

Impulse response functions

Figure 9 shows the impulse responses to a shock to $ENTSENT\_NEG$ (left) and $SENTNEG$ (right) in the panel VAR. As in the aggregate VAR, we see significant responses of implied and realized volatility (top row) to a shock to $ENTSENT\_NEG$. However, the response now peaks much sooner — at two months for implied volatility and at one month for realized volatility. In fact, if we omit time zero (the time of the initial shock, at which our variable ordering forces the volatility responses to be zero), the response of realized volatility is better described as declining rather than hump-shaped. These results suggest that the information in our news measures is absorbed more quickly at the company-specific level than at the aggregate level. The same pattern holds for the responses to $ENTSENT\_POS$ (available in the supplementary appendix), but we focus on negative news for brevity.

dimensions along which agents need to make an attention allocation decision.

[29] Comparing this to Model 8, we see that dropping $ARTPERC$ and r^- from this regression has very little effect on the remaining coefficient estimates (the latter observation is consistent with our finding from Footnote 22 that r^- has very little impact when added to our VARs).

Comparing the right panel of Figure 9 with the left panel we see that a shock to $SENTNEG$ does not produce a significant response in either implied or realized volatiity. In other words, consistent with what we saw in the aggregate case, it is the interaction of unusualness with negative sentiment that yields an increase in volatility.[30]

As we will see in Section 6.2, the shape of the impulse response functions can be partly explained through a model of attention-constrained investors. The model predicts a humped response if investors are more tightly constrained in their ability to process news, with tighter constraints leading to a later peak, and a monotonically declining response when investors are relatively unconstrained. The contrast between our aggregate and company-level impulse response functions therefore supports the view that investors are better able to process micro information than macro information.

Incremental R^2

To further investigate the contrast between aggregate and company-level news, we examine the relative importance of each of the forecasting variables in our regressions. We measure importance through the change in R^2 in (10) (or the corresponding regression with aggregate variables) when we remove either the implied volatilities, the realized volatilities, or all the text based measures from the right hand side of the equation, while leaving the other right hand side variables untouched. We refer to the resultant drop in R^2 as the *incremental* R^2s associated with implied volatility, realized volatility, and the text based measures respectively. Table 10 shows the incremental R^2 from the four versions of this regression (single name and macro results for forecasting implied and realized volatilities).

The last column of Table 10 shows the incremental R^2 from the news-based measures. In every case (single-name and macro, implied and realized), the F-test comparing restricted and unrestricted models shows that the news-based variables remain significant after the inclusion of lagged implied and realized volatility. The incremental R^2 for the news-based measures is substantially larger (3–3.6% compared with 0.5–0.7%) in the macro regressions than in the single-name regressions. In other words, more of the information content of the news-based measures is already reflected in the other variables at the single-name level. This pattern further supports

[30] If we switch the order of the variables $ENTSENT\_NEG$ and $SENTNEG$, as we did in Figure 6, the response to a shock in $SENTNEG$ becomes significant, but the response to a shock in $ENTSENT\_NEG$ remains significant, reinforcing the importance of the interaction. Those results are omitted for brevity, and are available in the supplementary appendix.

the view that market participants are better able to process the company-specific information than the aggregate information in news flow.

Comparing the *ivol* and *rvol* columns indicates that implied volatility contributes relatively more explanatory power at the aggregate level than at the single-name level, while the opposite holds for realized volatility. Combining these observations with the previous paragraph suggests the following intuitively plausible picture. Traders in the VIX extract all relevant information from the time series of realized volatility but do not fully incorporate the information in news articles relevant to aggregate volatility. Traders that follow individual companies extract more of the company-specific information in the news but devote less attention to the pricing information contained in the time series of company-specific realized volatility.[31]

This pattern is consistent with our conjecture that investor inattention is an underlying cause of the results in our paper. In the macro context, where investors would have a harder time keeping up with all news flow about the 50 companies in our sample that would be relevant for the macro economy, our news-based measures have a much larger contribution than they do at the single-name level, where investors – especially with regard to the well-followed large financial companies of our sample – are more likely to absorb a larger fraction of the relevant news flow. We intrepret this finding as supportive of the relative micro efficiency and macro inefficiency of markets.

6.2 Rational inattention and information constraints

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.[32] 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.

[31] Another possibility is that single-name implied volatility is more noisy than the VIX because of the relatively lower liquidity of single-name options. Such implicit measurement error can therefore account partially for the lower explanatory power of implied volatility in our single name regression. Because we look at only large financial companies, it is likely that this effect is small.

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

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.[33] For simplicity, we suppose that X_t follows a stationary AR(1) process,

$$X_{t+1} = \rho X_t + a u_{t+1}, \tag{11}$$

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

$$Y_t = \sum_{\ell=0}^{\infty} b_\ell u_{t-\ell} + \sum_{\ell=0}^{\infty} c_\ell \epsilon_{t-\ell}, \tag{12}$$

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.[34]

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

[33] 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.

[34] We omit the precise definition of the information constraint because it takes several steps to develop. In the 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 (12).

stochastic discount factor of the form[35] $\exp(\lambda u_{t+1} - \lambda^2/2)$, where u_{t+1} is the innovation to the macro state in (11). 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[36]

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

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 (monotonically decreasing) if agents' information processing constraint is sufficiently tight (loose). Figure 10 shows the impulse response of V_t to an innovation in X_t when agents are information constrained (blue line) and unconstrained (dashed, red line).

To map this observation to the impulse response functions in Sections 5.1 and 6.1, think of implied volatility (either aggregate or single-name) as the price of a contract that imperfectly hedges either macro or company specific risk: higher levels of the implied volatility are associated with market stress, so a contract that pays more in these states partly offsets a macro or micro risk. Interpret the variable $ENTSENT\_NEG$ as a proxy for the macro or micro state. The model we have outlined predicts that when the information constraint between $ENTSENT\_NEG$ and implied volatility is tight, the impulse response function should be humped, just as we saw in Section 5.1. The tighter the constraint, the more delayed the peak response. The information constraint faced by agents limits how quickly innovations to the underlying state get incorporated in implied volatility. Similarly, when agents are relatively information unconstrained, the impulse response is monotonically decreasing, as we saw for realized volatility in Section 6.1.

A more precise mapping between the model and our application should recognize that $ENTSENT\_NEG$ is itself at best a noisy observation of the underlying 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

[35] 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.

[36] 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).

to u_{t+1}. In this interpretation, the impulse response functions we observe in Sections 5.1 and 6.1 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.

7 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 and positive news forecast 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:

- Our interacted measures remain significant when we control for other predictors of single-name volatility (lagged volatility and negative returns), indicating that the information in these news measures is not fully reflected in contemporaneous option prices or realized volatility. In panels that include the interacted and non-interacted news measures, only the interacted news measures are economically and statistically significant.

- 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 (positive) news predicts an increase (decrease) in implied and realized volatility over several months. The effect is stronger for our interacted variable than for negative or positive sentiment alone.

When comparing our single name and macro results, we show that our news measures have more incremental information content at the macro than the micro level. This suggests that market participants have a harder time incorporating all relevant news at the macro than the micro level. Furthermore, we find that news shocks affect realized and implied volatilities in a hump-shaped manner over time, with the peak occurring later at the macro level than the micro. A humped response would not obtain simply from the persistence of volatility: in this case the effect of a news shock would dissipate monotonically. The pattern of responses we find indicates that news is not absorbed by the market instantaneously, and the absorption is slower at the aggregate level than at the company-specific level. We argue that this type of response is consistent with investors who face constraints on the rate at which they can process information and who process micro information more easily than macro 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 Details of entropy calculation

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 $m_i(t)$ in (4) to be

$$m_i(t) \equiv \frac{\tilde{c}(\{w_1 w_2 w_3 w_4\}) + 1}{\tilde{c}(\{w_1 w_2 w_3\}) + 4}.$$

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. 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 approximate $\tilde{c}(\{w_1 w_2 w_3 w_4\})$ in a given training window by only counting the occurrences of those 4-grams which are among the most frequently occurring 5,000 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'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. Indeed, across the 228 months in our sample, the maximum least-frequently-observed n-gram empirical probability is only 0.012%. Rerunning the analysis using the top 4,000 or 10,000 n-grams in each month leaves our results largely unchanged, suggesting the analysis isn't sensitive to this issue.

Our results are not very sensitive to the following modeling assumptions discussed here and in Section 2.3: setting unobserved m_i's to 0.25; having the rolling window for training be two years; the choice of three months for the training window offset; or the number of n-grams we use to estimate $\tilde{c}$.

A.2 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_`.[37] 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.

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

Our n-grams undergo a further data-cleaning step. 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.

A.3 Subperiod analysis

The results discussed in this section are not included in the paper to conserve space, but are available in the supplementary appendix.

Single name panels

We repeated the panel analysis of Section 4 over the pre-crisis time-period from June 1998 to December 2006. We estimated a modified version of equation (8) where we dropped the implied volatility data series (since they only start in January of 2005). We ran the panel using 2 and 3 lags, to generate tables comparable to Table 9. The results for $SENTPOS$ and $ENTPOS$ were weaker in the pre-crisis subsample (in fact, the coefficients on $ENTPOS$ and $ENSENT\_POS$ were positive). However the loadings on $ENTNEG$, $SENTNEG$ and $ENTSENT\_NEG$, while smaller in magnitude, were generally significant and had the appropriate (positive) sign. The stronger results evident in Table 9 suggest that our model was particularly effective in forecasting single name realized volatility during the financial crisis, but negative entropy and sentiment were significant both statistically and economically in forecasting realized volatility even in the pre-crisis period.

Aggregate impulse responses

Finally, we re-estimated the aggregate-level impulse response functions from Section 5.1 in the pre-crisis time period from April 1998 to December 2006. All aggregate level responses were quantitatively and qualitatively similar to the full-sample results, though the significance levels of the impulse responses were lower (which is at least partially attributable to the shorter data sample) and $SENTPOS$ tended to outperform $ENSENT\_POS$ ($ENTSENT\_NEG$ still outperformed $SENTNEG$). We conclude that the relationships among S&P realized volatility, the VIX and our aggregate level entropy and sentiment measures were largely the same in the pre- and post-crisis subsamples of the data.

A.4 Rational inattention

Proposition 3 of Maćkowiak and Wiederholt (2009a) shows that the optimal Y_t in (12) has

$$b_\ell = a\left(\rho^\ell - \frac{1}{2^{2\kappa}}\left(\frac{\rho}{2^{2\kappa}}\right)^\ell\right), \tag{13}$$

and

$$c_\ell = c_0\left(\frac{\rho}{2^{2\kappa}}\right)^\ell, \tag{14}$$

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 $c_0 = 0$ 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 (13)–(14) (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 (11) 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 (13), the response of $V_t, V_{t+1}, \dots$ to an impulse of $u_t = 1$ is given by $b_0\lambda + \rho b_t$, $t = 0, 1, \ldots$. As illustrated in Figure 10, 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, 2016, “Measuring economic policy uncertainty,” *Quarterly Journal of Economics* 131 (4), 1593–1636.

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.

Chernov, M., 2007, “On the role of risk premia in volatility forecasting,” *Journal of Business and Economic Statistics*, 25 (4), 411–426.

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.

Das, S., 2014, “Text and Context: Language Analytics in Finance,” *Foundations and Trends in Finance* 8(3), 145–261.

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, 2016, “The pitch rather than the pit: investor inattention, trading activity, and FIFA World Cup matches,” *Journal of Money, Credit and Banking*, forthcoming.

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

Feinstein, S., 1989, “The Black-Scholes formula is nearly linear in sigma for at-the-money options: Therefore implied volatilities from at-the-money options are virtually unbiased,” *Working Paper Federal Reserve Bank of Atlanta.*

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

Glasserman, P., and Mamaysky, H., 2016, “Market efficiency with micro and macro information,” working paper, Columbia Business School.

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.

Jiao, P., A. Veiga, and A. Walther, 2016, “Social media, news media and the stock market,” working paper.

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.

Kacperczyk, M. T., S. Van Nieuwerburgh, and L. Veldkamp, 2016, “A rational theory of mutual funds’ attention allocation,” *Econometrica*, 84 (2), 571–626.

Kogan, S., B. Routledge, J. Sagi, and N. Smith, 2011, “Information content of public firm disclosures and the Sarbanes-Oxley act,” working paper.

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.

Loughran, T., and McDonald, B., 2016, “Textual analysis in finance and accounting: A survey,” *Journal of Accounting Research*, 54 (4), 1187–1230.

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.

Poon, S.-H. and C. Granger, 2003, "Forecasting volatility in financial markets: A review," *Journal of Economic Literature*, 41, 478–539.

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.

Solomon, D.H., Soltes, E. and Sosyura, D., 2014, "Winners in the spotlight: Media coverage of fund holdings as a driver of flows," *Journal of Financial Economics* 113 (1), 53–72.

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.

Single name sentiment/entropy correlations with volatility

	ARTPERC	ENTNEG	ENTPOS	ENTALL	SENTNEG	SENTPOS
Mean corr with ivol	-0.004	0.197	-0.004	0.095	0.306	-0.097
SE ivol	0.036	0.026	0.019	0.024	0.024	0.017
Mean corr with rvol	0.085	0.208	-0.012	0.112	0.220	-0.074
SE rvol	0.026	0.019	0.018	0.016	0.019	0.014

Table 3: This table shows the average correlation between different entropy and sentiment measures and 1 month at-the-money implied or intramonth realized volatilities for the 50 stocks in our sample. Also shown is the average correlation between the percent of all monthly articles about a given company ($ARTPERC$) and realized and implied volatilities. 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. The realized volatility is available for all stocks. 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.

	Mean	Min	Max	SD
ENTNEG	7.401	2.140	11.592	1.837
ENTPOS	6.443	2.172	16.747	2.092
ENTALL	7.446	4.724	9.908	1.066
SENTNEG	3.744	1.484	8.077	1.265
SENTPOS	2.233	0.819	4.958	0.594
ENTSENT_NEG	28.156	7.881	77.348	13.948
ENTSENT_POS	14.189	4.947	45.117	5.711
VIX	21.444	10.420	59.890	8.272
SPX_rvol	17.675	4.140	87.880	10.776

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. $ENTSENT\_NEG$ ($ENTSENT\_POS$) interacts $SENTNEG$ ($SENTPOS$) with $ENTNEG$ ($ENTPOS$). All data series are monthly, and run from April 1998 to December 2014.

	SENTNEG	SENTPOS	ENTALL	ENTNEG	ENTPOS	ENTSENT_NEG	ENTSENT_POS	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		
ENTSENT_POS	-0.15	0.54	0.16	-0.03	0.73	-0.14	1.00	
VIX	0.46	-0.37	0.30	0.48	0.15	0.60	-0.14	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. $ENTSENT\_NEG$ ($ENTSENT\_POS$) interacts $SENTNEG$ ($SENTPOS$) with $ENTNEG$ ($ENTPOS$).

	ARTPERC	SENTPOS	ENTPOS	ENTSENT_POS	SENTNEG	ENTNEG	ENTSENT_NEG
ivol_l1	0.001	-0.001	0.001	0.000	0.001	0.001	0.001
ivol_l2	0.003***	-0.001	-0.001	-0.001	0.000	-0.003**	-0.002
rvol_l1	0.000	0.000	0.001	0.001	0.004***	0.004***	0.006***
rvol_l2	-0.002***	0.000	0.000	0.000	-0.001	0.001	0.000
ret_mi_l1	0.007***	-0.007*	-0.007	-0.007	0.003	0.009**	0.007
ret_mi_l2	-0.001	0.000	-0.005	-0.005	0.003	0.006	0.007
ARTPERC_l1	0.367***						
ARTPERC_l2	0.217***						
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***
Sum NM 2	0.583***	0.241***	0.29***	0.207***	0.39***	0.346***	0.307***
p-val NM	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
R2 adj	0.264	0.042	0.052	0.028	0.154	0.136	0.166

Table 7: This table reports the results of the panel model from (9). The dependent variable is shown in the column heading, with the regressors in the rows. The notation _lN indicates the variable is lagged by N months. The row labeled “Sum NM 2” 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. The regressions include data from Mar 2005 through Dec 2014.

	ivol (1)	ivol (2)	ivol (3)	ivol (4)	ivol (5)	ivol (6)	ivol (7)	ivol (8)	ivol (9)
ivol_l1	0.289***	0.307***	0.305***	0.305***	0.304***	0.252***	0.251***	0.243**	0.283***
ivol_l2	0.067	0.094	0.080	0.080	0.091	0.103	0.102	0.094	0.064
rvol_l1	0.222***	0.209***	0.209***	0.208***	0.207***	0.226***	0.222***	0.219***	0.246***
rvol_l2	0.119***	0.102***	0.108**	0.107**	0.098**	0.104**	0.100**	0.103**	0.086
ret_mi_l1	0.371***	0.364***	0.375***	0.374***	0.360***	0.420***	0.411***	0.400***	
ret_mi_l2	-0.165*	-0.174*	-0.165	-0.166	-0.172*	-0.173*	-0.174*	-0.163	
ARTPERC_l1	-0.681	-0.677	-0.666	-0.638	-0.732	-0.806	-0.913*	-0.670	
ARTPERC_l2	0.224	0.270	0.148	0.168	0.233	0.149	0.247	0.108	
SENTPOS_l1		-0.422						0.035	0.044
SENTPOS_l2		-0.235						0.127	0.181
ENTPOS_l1			-0.369						
ENTPOS_l2			-0.395						
ENTSENT_POS_l1				-0.635				-0.715	-0.778
ENTSENT_POS_l2				-0.406				-0.567	-0.654
SENTNEG_l1					0.802**			0.773	0.953
SENTNEG_l2					0.467			-0.473	-0.497
ENTNEG_l1						0.225			
ENTNEG_l2						0.757**			
ENTSENT_NEG_l1							0.739**	0.310	0.248
ENTSENT_NEG_l2							0.807**	1.247**	1.264**
Sum NM 2	-0.456	-0.657*	-0.764	-1.041**	1.27***	0.982*	1.546***	-1.282*	-1.432*
p-val NM	[0.378]	[0.084]	[0.145]	[0.031]	[0.006]	[0.070]	[0.001]	[0.076]	[0.055]
R2 adj	0.574	0.576	0.576	0.576	0.577	0.593	0.594	0.582	0.574
Text F-test	-	[0.205]	[0.346]	[0.098]	[0.017]	[0.116]	[0.006]	[0.016]	[0.008]

Table 8: This table reports the results of the panel model from (8). The dependent variable is shown in the column heading, with the regressors in the rows. The notation _lN indicates the variable is lagged by N months. The row labeled "Sum NM 2" shows the sum of the two bottom-most coefficients in each column, or the sum of coefficients of $ENTSENT\_POS$ where multiple news measures are present. 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. "Text F-test" is the p-value from an F-test that restricts all text-based measures (excluding $ARTPERC$) in the regression to be zero. The regressions include data from Mar 2005 through Dec 2014.

	rvol (1)	rvol (2)	rvol (3)	rvol (4)	rvol (5)	rvol (6)	rvol (7)	rvol (8)	rvol (9)
ivol_l1	0.245***	0.268***	0.265***	0.264***	0.263***	0.295***	0.292***	0.281***	0.368***
ivol_l2	0.091**	0.120***	0.128***	0.128***	0.118**	0.118**	0.118**	0.124**	0.069
rvol_l1	0.349***	0.331***	0.312***	0.311***	0.329***	0.311***	0.306***	0.292***	0.364***
rvol_l2	0.140***	0.122**	0.130**	0.130**	0.116**	0.116**	0.108**	0.110**	0.082
ret_mi_l1	0.865***	0.859***	0.928***	0.925***	0.850***	0.860***	0.845***	0.887***	
ret_mi_l2	-0.077	-0.092	-0.108	-0.113	-0.089	-0.105	-0.104	-0.119	
ARTPERC_l1	-1.269**	-1.246**	-1.049*	-0.973*	-1.314**	-1.321**	-1.432**	-1.277**	
ARTPERC_l2	-0.024	0.026	-0.244	-0.254	-0.048	-0.146	-0.031	-0.241	
SENTPOS_l1		-0.684*						-0.210	-0.166
SENTPOS_l2		-0.401						0.183	0.336
ENTPOS_l1			0.052						
ENTPOS_l2			-0.431						
ENTSENT_POS_l1				-0.556				-0.469	-0.661
ENTSENT_POS_l2				-0.662				-0.925	-1.182
SENTNEG_l1					1.646***			0.527	0.800
SENTNEG_l2					0.234			-0.674	-0.636
ENTNEG_l1						0.985*			
ENTNEG_l2						0.521			
ENTSENT_NEG_l1							1.977***	1.893**	1.887**
ENTSENT_NEG_l2							0.451	1.023	0.979
Sum NM 2	-1.293**	-1.085**	-0.379	-1.217*	1.88***	1.506*	2.429***	-1.394	-1.843
p-val NM	[0.027]	[0.023]	[0.628]	[0.068]	[0.004]	[0.074]	[0.001]	[0.230]	[0.129]
R2 adj	0.587	0.589	0.583	0.584	0.591	0.592	0.595	0.588	0.568
Text F-test	-	[0.064]	[0.689]	[0.186]	[0.002]	[0.162]	[0.000]	[0.016]	[0.006]

Table 9: This table reports the results of the panel model from (8). The dependent variable is shown in the column heading, with the regressors in the rows. The notation $\_lN$ indicates the variable is lagged by N months. The row labeled "Sum NM 2" shows the sum of the two bottom-most coefficients in each column, or the sum of coefficients of $ENTSENT\_POS$ where multiple news measures are present. 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. "Text F-test" is the p-value from an F-test that restricts all text-based measures (excluding $ARTPERC$) in the regression to be zero. The regressions include data from Mar 2005 through Dec 2014.

Text based indicator contribution to volatility forecasting

	LHS	full_R2	start	end	ivol	rvol	news
ivol s-n	ivol	0.582	Mar 2005	Dec 2014	0.027	0.041	0.005
ivol s-n F-test					0.007	0.000	0.008
rvol s-n	rvol	0.576	Mar 2005	Dec 2014	0.025	0.045	0.007
rvol s-n F-test					0.000	0.000	0.006
ivol macro	ivol	0.757	Jun 1998	Dec 2014	0.094	0.004	0.030
ivol macro F-test					0.000	0.235	0.000
rvol macro	rvol	0.615	Jun 1998	Dec 2014	0.071	0.012	0.036
rvol macro F-test					0.000	0.077	0.048

Table 10: The *Full model* column gives unadjusted R^2's from the regression in (10). The columns labeled *ivol*, *rvol*, and *news* show the drop in the full model unadjusted R^2 if $ivol$, $rvol$ or the 4 news measures are removed from the right side of equation (10). The rows labeled "F-test" show the p-value of an F-test comparing the restricted to the unrestricted model (the standard errors are clustered by time for the panel regressions, and Newey-West with auto-lag selection is used for the macro tests). The top half of the table refers to the single name panel regressions for implied and realized volatilities with the full model coefficient estimates shown in Tables 8 and 9. The lower half of the table shows the results of the VIX (here called *ivol macro*) and the SPX_RVOL (*rvol macro*) regressions from the VARs discussed in Section 5.

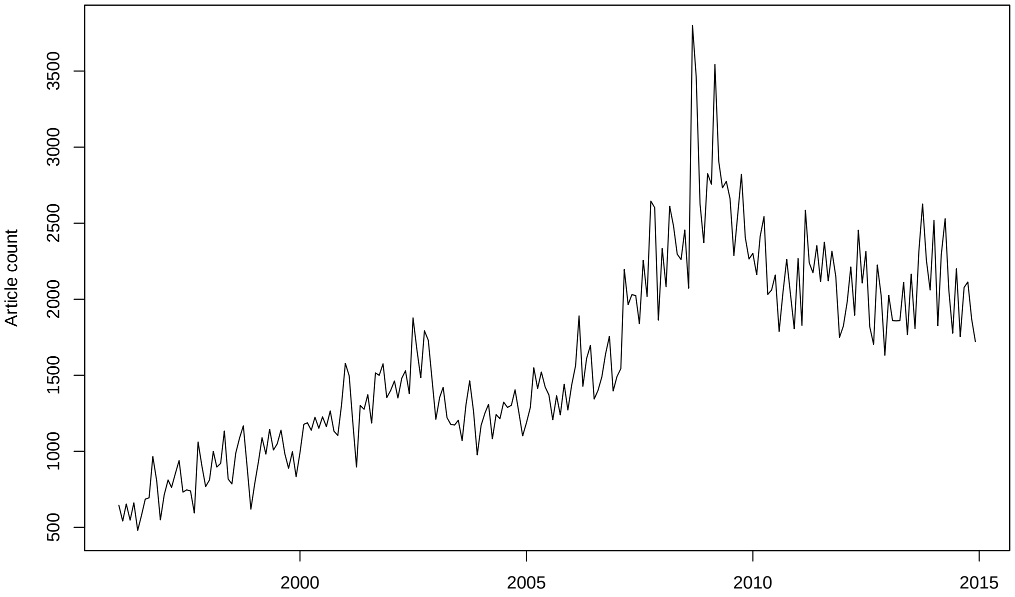

Figure 1: Monthly article count in the Thomson Reuters news sample.

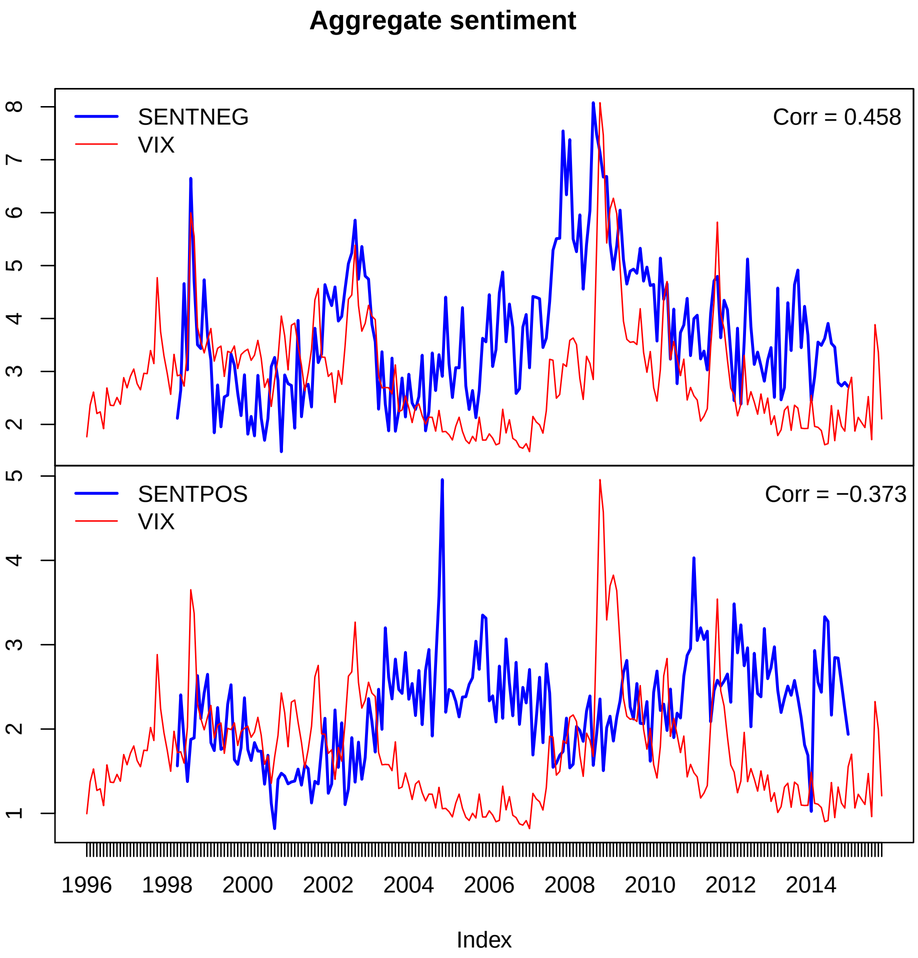

Figure 2: Monthly plots of $SENTNEG(t)$ and $SENTPOS(t)$ as defined in equation (7). Each series computes the proportion of all n-grams in a given month that are classified as having either positive or negative sentiment. Superimposed on each sentiment series is the scaled VIX index. Correlation between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

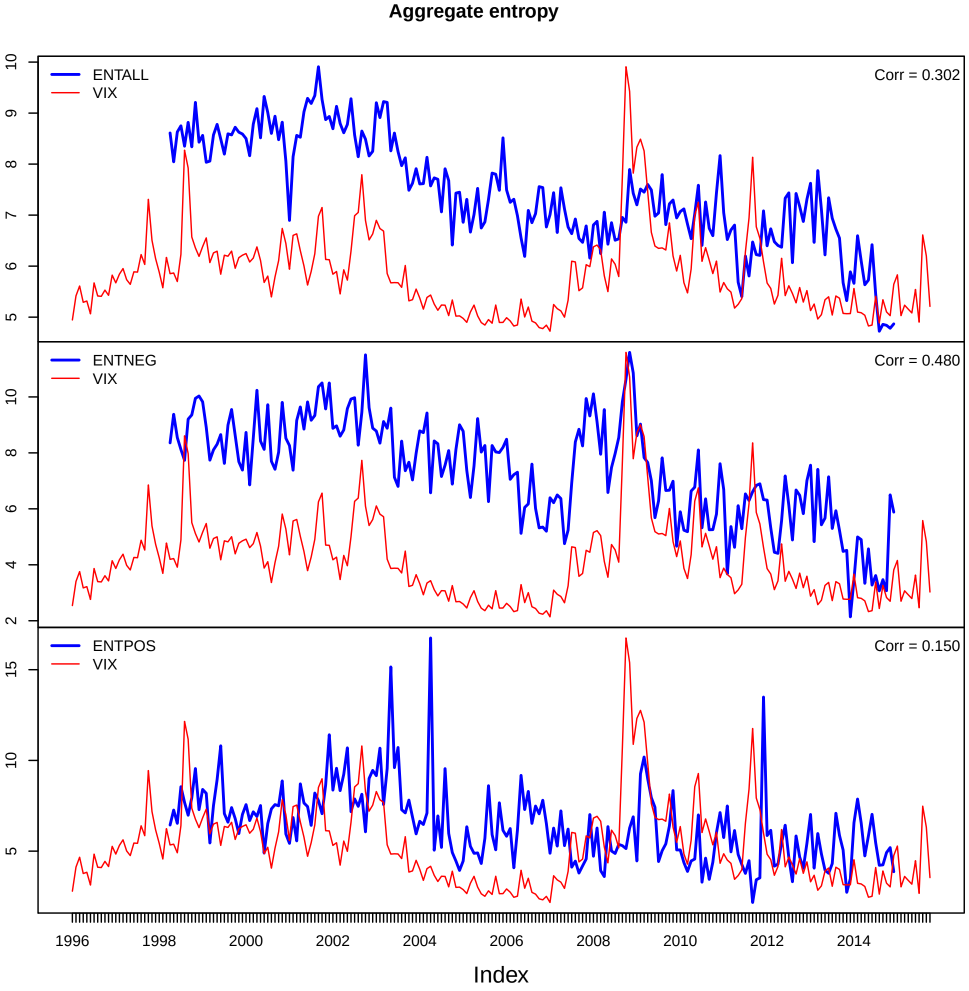

Figure 3: Monthly plots of $ENTALL(t)$, $ENTNEG(t)$ and $ENTPOS(t)$ as defined in Section 2.3. Each series is the first principal component of the associated single name entropy measures, for those names with observations available in all time periods of the sample. Superimposed on each entropy series is the scaled VIX index. Correlation between entropy and VIX is shown in the upper right hand corner of each chart.

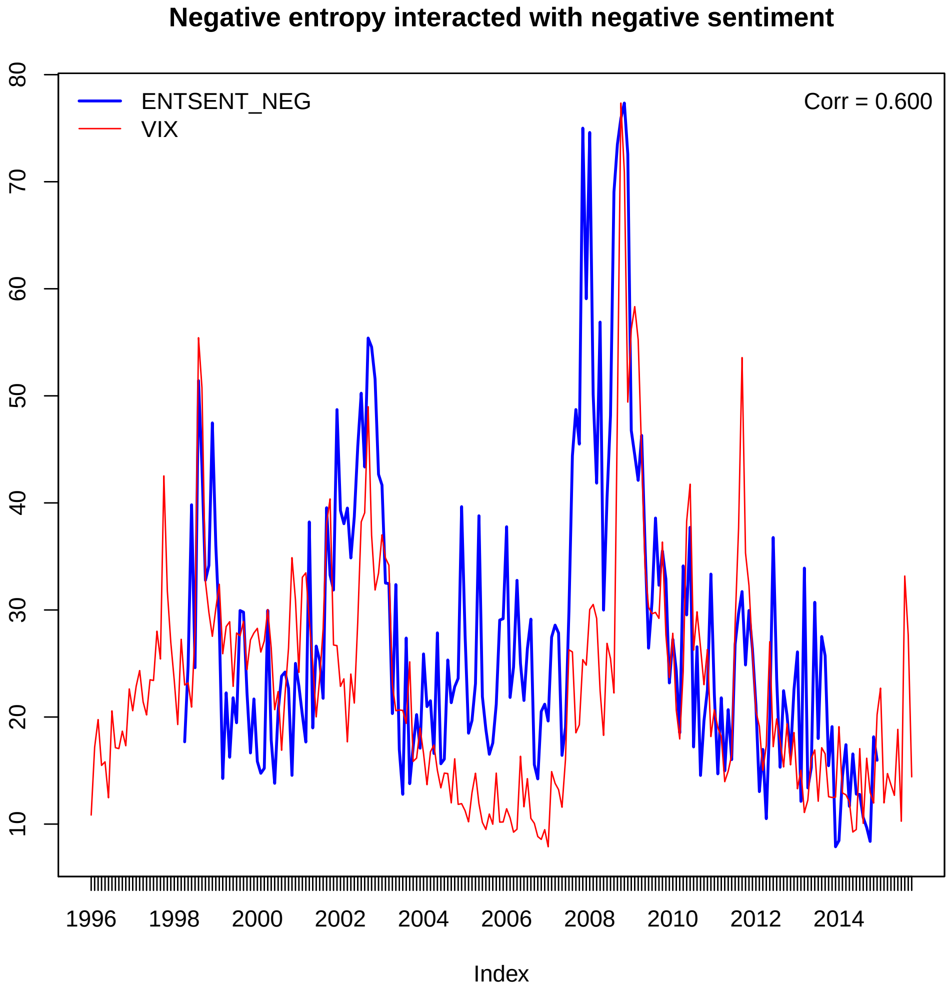

Figure 4: Monthly plot of $ENTSENT\_NEG(t) \equiv ENTNEG(t) \times SENTNEG(t)$. The entropy series is the first principal component of the associated single name entropy measures, for those names with observations available in all time periods of the sample. $SENTNEG$ is defined in (7). Superimposed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

Aggregate VAR – Impulse responses to negative news

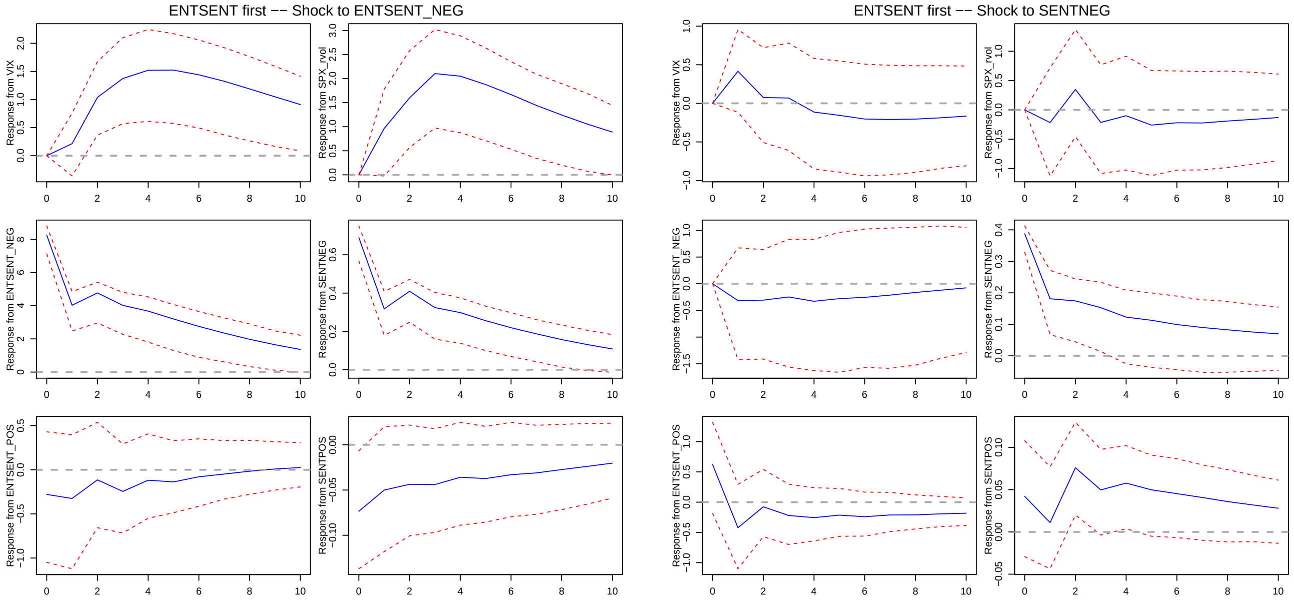

Figure 5: Impulse response functions for a shock to $ENTSENT\_NEG$ (left) and $SENTNEG$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

Aggregate VAR (order reversed) – Impulse responses to negative news

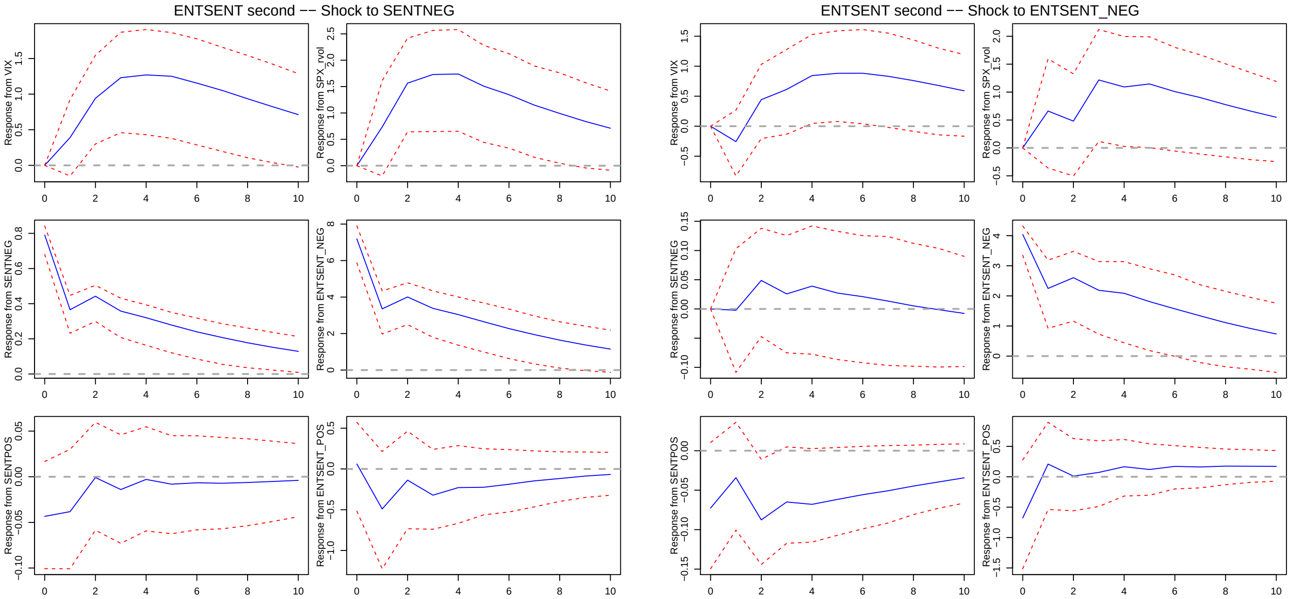

Figure 6: Impulse response functions for a shock to $SENTNEG$ (left) and $ENTSENT\_NEG$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

Aggregate VAR – Impulse responses to positive news

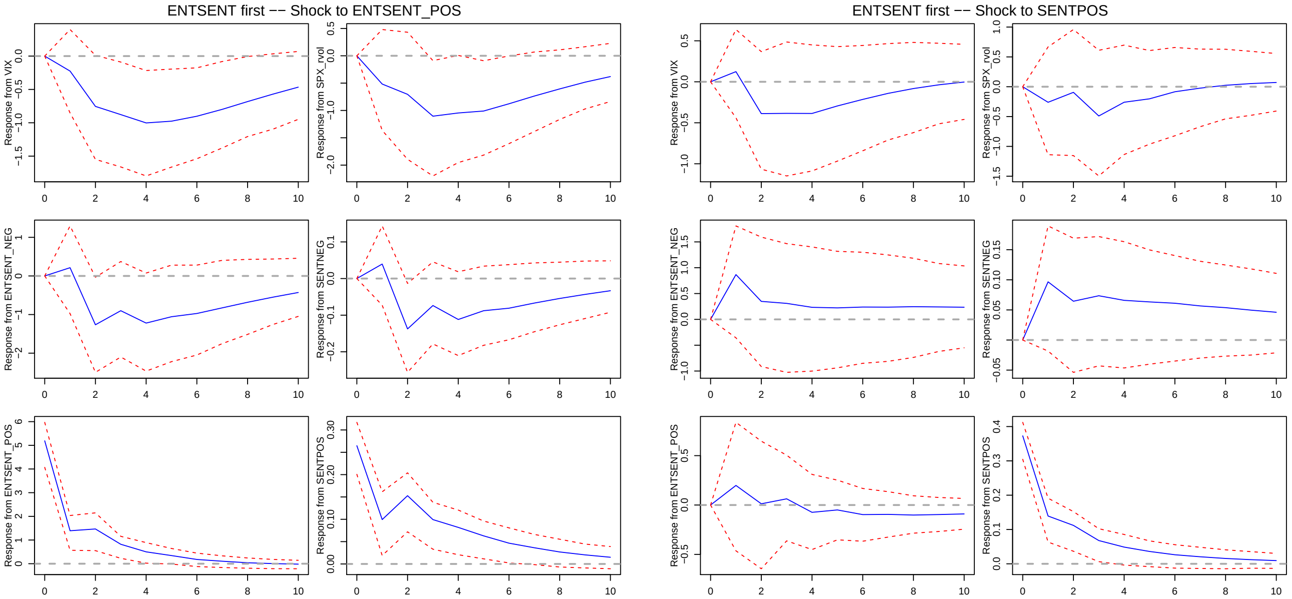

Figure 7: Impulse response functions for a shock to $ENTSENT\_POS$ (left) and $SENTPOS$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

Aggregate VAR (order reversed) – Impulse responses to positive news

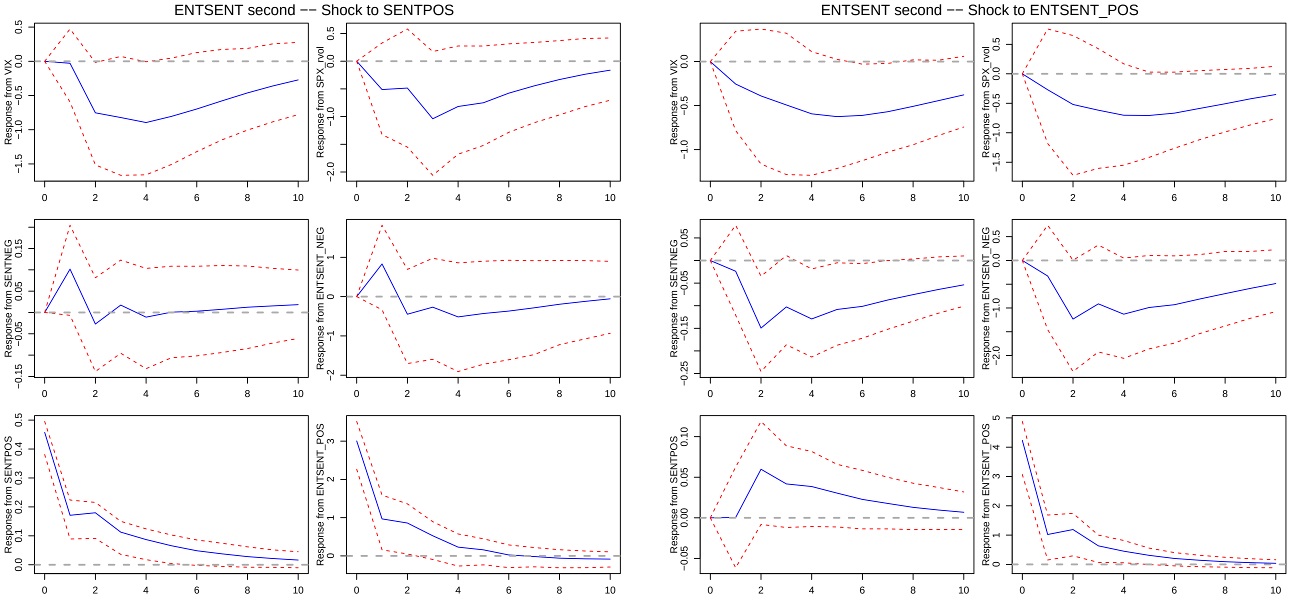

Figure 8: Impulse response functions for a shock to $SENTPOS$ (left) and $ENTSENT\_POS$ (right). The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

Company-level panel VAR – Impulse responses to negative news

ENTSENT_NEG shock

SENTNEG shock

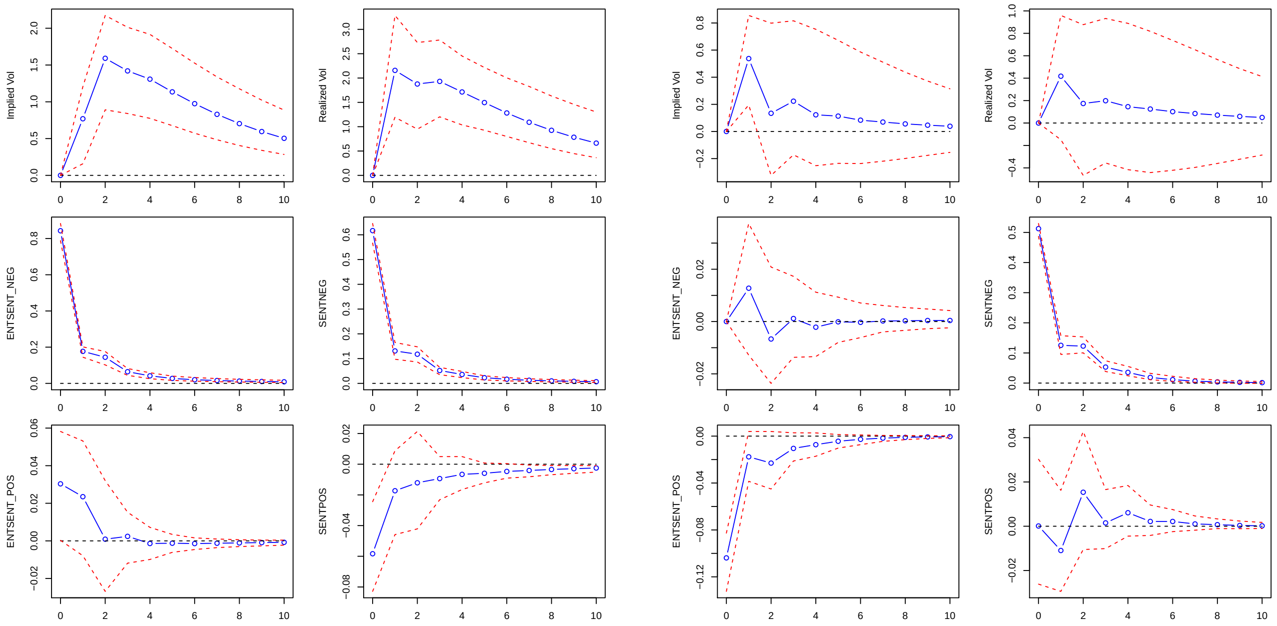

Figure 9: Impulse response functions for a shock to $ENTSENT\_NEG$ (left) and $SENTNEG$ (right) in the company-level panel VAR. The order of the variables in the VAR model matches the order of the figures in each block of six, reading left to right, then top to bottom. Dashed lines show 95 percent bootstrap confidence intervals. The horizontal time axis is in months.

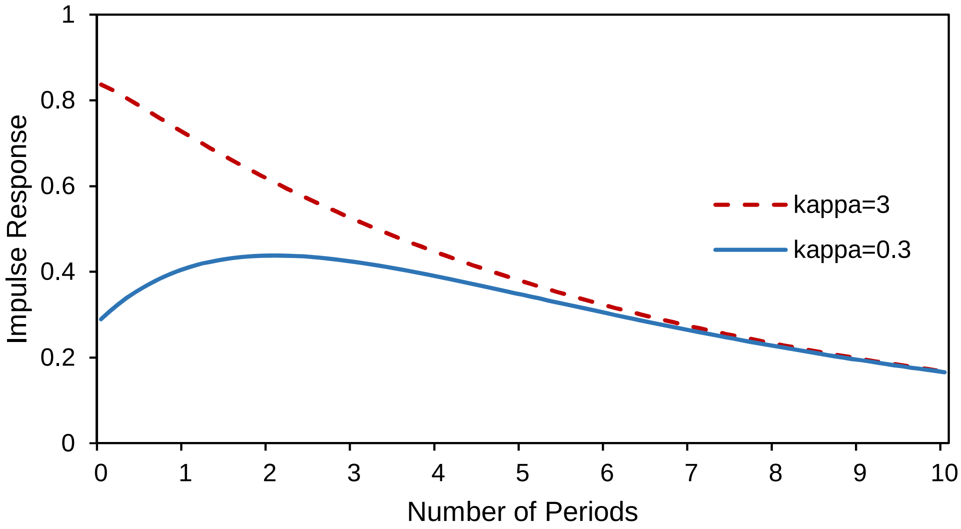

Figure 10: This figure shows impulse response functions in the model of Section 6.2. The response is hump-shaped for small κ (a tight information constraint) and monotonically decreasing for large κ. The other parameters are $\rho = 0.85$, $\lambda = 0$, and $a = 1$.