Does Unusual News Forecast Market Stress?

Harry Mamaysky and Paul Glasserman
Columbia University

6th Annual High Frequency Finance and Analytics Conference
October 28, 2015
Introduction

- Automated processing of natural language is opening a previously unavailable window into market behavior

  - Unavailable to both practitioners and academics!

- Prior work has documented intriguing relationships between short-term price responses and news sentiment

- This area of research has the potential to fundamentally change finance practice
Contributions

- Study volatility rather than directional moves
- A novel method for extracting content from text – *unusualness*
  - The information content of unusualness interacted with sentiment dominates that of sentiment
  - Effects occur over months
- At a single-name level unusualness forecasts future implied volatility
  - For the largest international financial firms, such information is largely incorporated into prices
- At the aggregate level unusualness forecasts future implied and realized volatility, even after controlling for current market information
- A simple model with rational inattention provides one theoretical explanation for our aggregate level results
An example

- Two phrases from news articles from September 2008:
  - “the collapse of Lehman”
  - “problem accessing the internet”

- Both contain negative words

- Intuitively, the first one seems to have more meaningful content

- According to our definition of unusualness, the first is one of the most unusual negative phrases in September 2008, and the second one of the least unusual

- How to come up with a probabilistic model for language?
Counting positive and negative words matters for short-term aggregate and single-name price responses
  - Tetlock (2007), Tetlock et al. (2008), Garcia (2013)
Use market responses to determine importance of words
Finance-specific word lists
  - Louhgran and McDonald (2011)
Institutional order flow ahead of news
  - Hendershott et al. (2014)
Central bank tone
  - Schmeling and Wagner (2015)
Newness of information
  - Tetlock (2011)
N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Consider word sequence

\[ w_1^N \equiv w_1 w_2 \cdots w_N \]

Using chain rule, we can write

\[
P(w_1^N) = P(w_1)P(w_2|w_1)P(w_3|w_1^2) \cdots P(w_N|w_1^{N-1}) \tag{1}
\]

For a 4-gram model, we assume

\[
P(\text{industry}|\ldots \text{ deal comes amid sweeping changes in the US cable}) = P(\text{industry}|\text{the US cable})
\]

Dropping the first few terms in (1), we can then write

\[
P(w_1^N) \approx \prod_{k=4}^{N} P(w_k|w_{k-3}w_{k-2}w_{k-1})
\]
N-grams

Likelihood of a corpus

Using a training corpus, we estimate the 4-gram probability of word $k$:

$$m(w_k | w_{k-3} w_{k-2} w_{k-1}) = \frac{C(w_{k-3} w_{k-2} w_{k-1} w_k)}{C(w_{k-3} w_{k-2} w_{k-1})}$$

where $C$ is the count operator

In a new $N$-word corpus, we compute its per word log probability score $1/N \log P(w_1^N)$:

$$\frac{1}{N} \sum_{k=4}^N \log m(w_k | w_{k-1}^{k-3}) = \sum_{W \in \text{All 4-grams}} p(W) \log m(w_4 | w_1 w_2 w_3)$$

where $p$ is the in-sample probability of the 4-gram $W = w_1 w_2 w_3 w_4$

The negative of the above converges to the cross-entropy of $m$ (model) with respect to the true word sequence probabilities
Entropy

- Entropy for a given collection of 4-grams in month $t$ is given by

$$-p_t \cdot \log m_t$$

- $m_t$ comes from training corpus with $s \in t - 27, \ldots, t - 4$
- $p_t$ comes from some subset of current data
- Subsets and entropies of interest:

<table>
<thead>
<tr>
<th>Subset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTALL</td>
<td>all month $t$ n-grams or n-grams from articles mentioning a specific company</td>
</tr>
<tr>
<td>ENTNEG</td>
<td>n-grams with negative (and no positive) words</td>
</tr>
<tr>
<td>ENTUNC</td>
<td>n-grams with uncertain words</td>
</tr>
<tr>
<td>ENTPOS</td>
<td>n-grams with positive (and no negative) words</td>
</tr>
</tbody>
</table>

- Adjustment for never-seen n-grams:

$$m_i(t) \equiv \frac{C(\omega_1\omega_2\omega_3\omega_4) + 1}{C(\omega_1\omega_2\omega_3) + 4}$$
Sentiment

- Examples of sentiment words:
  - **NEGATIVE**: closed, fears, disrupting, losses, accused
  - **UNCERTAIN**: almost, approximately, assumed, contingent, believed
  - **POSITIVE**: leading, gained, strengthened, boosted, prosperity

- Likewise we can classify 4-grams by sentiment, for the entire sample in month $t$ or those n-grams coming from articles about a given company

<table>
<thead>
<tr>
<th>$\text{SENTNEG}$</th>
<th>fraction of n-grams with negative (and no positive) words</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SENTUNC}$</td>
<td>fraction of n-grams with uncertain words</td>
</tr>
<tr>
<td>$\text{SENTPOS}$</td>
<td>fraction of n-grams with positive (and no negative) words</td>
</tr>
</tbody>
</table>
Example time $t$ corpus

<table>
<thead>
<tr>
<th>Dog Article</th>
<th>Cat Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) the dog goes home</td>
<td>(3) the cat is always</td>
</tr>
<tr>
<td>(2) the dog is happy</td>
<td>(4) cat is always angry</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{ENTALL}_{\text{dog}} &= -\frac{1}{2} \log(m_1) - \frac{1}{2} \log(m_2) \\
\text{ENTNEG}_{\text{cat}} &= -\log(m_4) \\
\text{ENTPOS}_{\text{dog}} &= -\log(m_2) \\
\text{SENTPOS}_{\text{dog}} &= \frac{1}{2} \\
\text{SENTNEG}_{\text{cat}} &= \frac{1}{2} \\
\text{SENTPOS}_{\text{cat}} &= \frac{1}{4} \\
\text{SENTNEG}_{\text{dog}} &= \frac{1}{4}
\end{align*}
\]

\[
\begin{align*}
\text{ENTALL} &= -\sum_{1=1}^{4} \frac{1}{4} \log(m_i) \\
\text{ENTPOS} &= -\log(m_2) \\
\text{ENTNEG} &= -\log(m_4) \\
\text{SENTNEG} &= \frac{1}{4} \\
\text{SENTPOS} &= \frac{1}{4}
\end{align*}
\]
## N-grams

### Notation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comment</th>
<th>Single-name</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTNEG</td>
<td>n-grams with negative words</td>
<td>✓</td>
<td>First PC of single name measures</td>
</tr>
<tr>
<td>ENTUNC</td>
<td>n-grams with uncertain words</td>
<td>✓</td>
<td>– same –</td>
</tr>
<tr>
<td>ENTPOS</td>
<td>n-grams with positive words</td>
<td>✓</td>
<td>– same –</td>
</tr>
<tr>
<td>ENTALL</td>
<td>All n-grams</td>
<td>✓</td>
<td>– same –</td>
</tr>
<tr>
<td>SENTNEG</td>
<td>Contain negative words</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SENTUNC</td>
<td>Contain uncertain words</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SENTPOS</td>
<td>Contain positive words</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ENTSENT_NEG</td>
<td>ENTNEG $\times$ SENTNEG</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ENTSENT_POS</td>
<td>ENTPOS $\times$ SENTPOS</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Empirical implementation

Data

- All articles from Thomson-Reuters archive for 50 largest global financial firms (as of February 6, 2015)
- Companies include: Berkshire Hathaway, Wells Fargo, Ind & Comm Bank of China, JP Morgan, HSBC, etc.
- Market caps range from US$ 40bln to US$ 300+bln
- Data from 1996 to 2014 for 228 year-months
- 367,331 articles
- Average of 1,611 articles per month
- Sentiment dictionary from Loughran-McDonald (http://www3.nd.edu/~mcdonald/Word_Lists.html)
Empirical implementation

Article count per month

![Article count over time graph](image-url)
## Data summary by country

<table>
<thead>
<tr>
<th>Country</th>
<th>Avg mkt cap (usd)</th>
<th>Percent of all articles</th>
<th>Number of firms</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNITED STATES</td>
<td>127.07</td>
<td>44.25</td>
<td>15</td>
</tr>
<tr>
<td>BRITAIN</td>
<td>80.01</td>
<td>19.11</td>
<td>5</td>
</tr>
<tr>
<td>AUSTRALIA</td>
<td>65.79</td>
<td>6.35</td>
<td>4</td>
</tr>
<tr>
<td>CANADA</td>
<td>68.16</td>
<td>6.08</td>
<td>3</td>
</tr>
<tr>
<td>SPAIN</td>
<td>64.62</td>
<td>4.68</td>
<td>2</td>
</tr>
<tr>
<td>FRANCE</td>
<td>66.05</td>
<td>4.63</td>
<td>2</td>
</tr>
<tr>
<td>NETHERLANDS</td>
<td>54.68</td>
<td>3.19</td>
<td>1</td>
</tr>
<tr>
<td>CHINA</td>
<td>129.89</td>
<td>2.70</td>
<td>8</td>
</tr>
<tr>
<td>GERMANY</td>
<td>71.56</td>
<td>2.22</td>
<td>1</td>
</tr>
<tr>
<td>SWITZERLAND</td>
<td>54.00</td>
<td>1.95</td>
<td>2</td>
</tr>
<tr>
<td>JAPAN</td>
<td>68.91</td>
<td>1.69</td>
<td>2</td>
</tr>
<tr>
<td>IRELAND</td>
<td>43.85</td>
<td>1.04</td>
<td>1</td>
</tr>
<tr>
<td>ITALY</td>
<td>58.89</td>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>BRAZIL</td>
<td>33.55</td>
<td>0.68</td>
<td>2</td>
</tr>
<tr>
<td>SWEDEN</td>
<td>45.03</td>
<td>0.63</td>
<td>1</td>
</tr>
</tbody>
</table>
### Examples of n-grams

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

**Table:** 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 appeared in that month, and the “Rank” column gives its rank by entropy contribution.
From Reuters “JPMorgan’s Dimon loses clout as reform critic” (May 11, 2012):

*But the revelation of a shocking trading loss of at least $2 billion from a failed hedging strategy diminishes Dimon’s credibility, and is already unleashing calls to get tougher on big banks.*
Single-name summary

<table>
<thead>
<tr>
<th></th>
<th>ENTNEG</th>
<th>ENTPoS</th>
<th>ENTALL</th>
<th>SENTNEG</th>
<th>SENTPOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean correlation</td>
<td>0.197</td>
<td>-0.003</td>
<td>0.095</td>
<td>0.309</td>
<td>-0.102</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.026</td>
<td>0.019</td>
<td>0.024</td>
<td>0.024</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Table: Average cross-sectional correlations of news-based measures with single-name implied volatilities (or VIX when unavailable).
Empirical implementation

Entropy and sentiment

Aggregate measures

NOTE: Scaled VIX shown in red
### Aggregate contemporaneous correlations

<table>
<thead>
<tr>
<th></th>
<th>SENTNEG</th>
<th>SENTPOS</th>
<th>ENTALL</th>
<th>ENTNEG</th>
<th>ENTPOS</th>
<th>ENTSENT_NEG</th>
<th>VIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>SENTNEG</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENTPOS</td>
<td>-0.14</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTALL</td>
<td>-0.18</td>
<td>-0.42</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTNEG</td>
<td>0.19</td>
<td>-0.44</td>
<td>0.71</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTPOS</td>
<td>-0.09</td>
<td>-0.16</td>
<td>0.56</td>
<td>0.34</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENTSENT_NEG</td>
<td>0.86</td>
<td>-0.32</td>
<td>0.19</td>
<td>0.64</td>
<td>0.08</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>VIX</td>
<td>0.46</td>
<td>-0.37</td>
<td>0.30</td>
<td>0.48</td>
<td>0.15</td>
<td>0.60</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Table:** 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 respectively. **ENTSENT_NEG** interacts **SENTNEG** with **ENTNEG**.
Single-name responses

N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Does unusualness matter for single names?

Two different questions

1. Does \textit{our measure} of unusualness matter for future implied volatility?
   
   ▶ How does this compare to sentiment-based measures?

2. Does the market fully incorporate this news-based information?
A comparison of news-based measures

For each single name $i$, run time series regression:

$$i\text{Vol}_{1mo}^i(t) = c + \sum_{l=1}^{L} s_l^i \text{NEWS\_MEASURE}^i(t - l) + \cdots + \epsilon^i(t)$$

- All NEWS\_MEASURE’s normalized to mean zero, unit variance
- Report cross-sectional averages of coefficients

$$s_l \equiv \frac{1}{N} \sum_{i=1}^{N} s_l^i$$

- Diagnostic is cross-sectional distribution of $R^2$’s compared to a baseline model
  - Baseline model: Explanatory variables are article and n-gram counts as percent of total, and their interaction.
Control regression with ARTICLE_PERCTOT

Market data regressed on lagged forecasting variables.
Names in cross-section = 41
Minimum observations = 60
Single-name responses

Results for **ENTNEG**

Reg summary for future implied vol on ENTNEG lags=6 ivol=false fwd step=0

Mean and 2xSE bands for cross-section of coefficients

1−CDF for R2

Present model
Article count benchmark
Diff AUC = 0.090

Market data regressed on lagged forecasting variables.
Names in cross-section = 38
Minimum observations = 60
# Single-name responses

## Ranking of news-based measures

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num Lags</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Lag ivol</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>Fwd Step</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENTNEG</td>
<td>0.0897***</td>
<td>0.0628***</td>
</tr>
<tr>
<td>ENTPOS</td>
<td>-0.052**</td>
<td>-0.0473**</td>
</tr>
<tr>
<td>ENTUNC</td>
<td>-0.0288</td>
<td>-0.0332*</td>
</tr>
<tr>
<td>ENTALL</td>
<td>-0.0278</td>
<td>-0.0312**</td>
</tr>
<tr>
<td>SENTNEG</td>
<td>0.139***</td>
<td>0.116***</td>
</tr>
<tr>
<td>SENTPOS</td>
<td>-0.0187</td>
<td>-0.0239*</td>
</tr>
<tr>
<td>SENTUNC</td>
<td>-0.0206</td>
<td>-0.0408***</td>
</tr>
<tr>
<td>ENTSENT_NEG</td>
<td>0.222***</td>
<td>0.178***</td>
</tr>
<tr>
<td>3 var Base mean R2</td>
<td>0.182***</td>
<td>0.145***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1 var Base mean R2</th>
<th>3 var Base mean R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.105</td>
<td>0.0795</td>
</tr>
<tr>
<td></td>
<td>0.28</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**Table:** The table reports the difference in average $R^2$ between the candidate NEWS_MEASURE specification and the control (either article or n-gram fractions).
Single-name responses

Results for **ENTNEG, SENTNEG, ENTSENT_NEG**

Market data regressed on lagged forecasting variables.
Names in cross-section = 38
Minimum observations = 60
Is the information already in prices?

- For each single name $i$, run time series regression:

$$iVol_{1mo}^i(t + f) = c + \sum_{l=1}^{L} v^i_l \cdot iVol_{1mo}^i(t - l)$$
$$+ \sum_{l=1}^{L} s^i_l \cdot NEWS\_MEASURE^i(t - l) + \cdots + \epsilon^i(t)$$

- Again look at cross-sectional means of the $s^i_l$'s

- Diagnostic is cross-sectional distribution of $R^2$'s compared to a baseline model
  - *Baseline model*: Explanatory variables are article and n-gram counts as percent of total, and their interaction.
Results for **ENTSENT_NEG**

Reg summary for future implied vol on ENTSENT_NEG lags=3 ivol=true fwd step=1

Mean and 2xSE bands for cross-section of coefficients

Market data regressed on lagged forecasting variables.
Names in cross-section = 38
Minimum observations = 60
Panel Regressions

Regress $IVol_t$ on lagged measures of volatility, negative returns, article percent counts, news measures

<table>
<thead>
<tr>
<th></th>
<th>ivol_l1</th>
<th>ivol_l2</th>
<th>rvol_l1</th>
<th>rvol_l2</th>
<th>ret_mi_l1</th>
<th>ret_mi_l2</th>
<th>ARTPERC_l1</th>
<th>ARTPERC_l2</th>
<th>SENTPOS_l1</th>
<th>SENTPOS_l2</th>
<th>ENTPOS_l1</th>
<th>ENTPOS_l2</th>
<th>ENTSENT_POS_l1</th>
<th>ENTSENT_POS_l2</th>
<th>SENTNEG_l1</th>
<th>SENTNEG_l2</th>
<th>ENTNEG_l1</th>
<th>ENTNEG_l2</th>
<th>ENTSENT_NEG_l1</th>
<th>ENTSENT_NEG_l2</th>
<th>Sum(N_M)</th>
<th>R2 adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>ivol</td>
<td>0.2900***</td>
<td>0.0457***</td>
<td>0.2727***</td>
<td>0.0896***</td>
<td>0.3943***</td>
<td>-0.1885***</td>
<td>-0.6682**</td>
<td>0.1601</td>
<td>-0.4271**</td>
<td>-0.1909</td>
<td>-0.4031</td>
<td>-0.3876</td>
<td>-0.6646***</td>
<td>-0.3734</td>
<td>0.7661***</td>
<td>0.4162</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-0.618**</td>
<td>0.582</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2867***</td>
<td>0.0287</td>
<td>0.2731***</td>
<td>0.0982***</td>
<td>0.4080***</td>
<td>-0.1780***</td>
<td>-0.6389*</td>
<td>0.0211</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.0211</td>
<td>0.0211</td>
<td>-0.3734</td>
<td>-0.6646***</td>
<td>-1.043***</td>
<td>-1.189***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-0.791**</td>
<td>0.583</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2860***</td>
<td>0.0286</td>
<td>0.2710***</td>
<td>0.0986***</td>
<td>0.4074***</td>
<td>-0.1783***</td>
<td>-0.6114*</td>
<td>0.0414</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.0414</td>
<td>0.0414</td>
<td>-1.043***</td>
<td>-0.791**</td>
<td>-1.043***</td>
<td>-1.189***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-1.043**</td>
<td>0.584</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2873***</td>
<td>0.0436**</td>
<td>0.2711***</td>
<td>0.0860***</td>
<td>0.3887***</td>
<td>-0.1880***</td>
<td>-0.7201**</td>
<td>0.1251</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.1251</td>
<td>0.1251</td>
<td>-1.043***</td>
<td>-1.043***</td>
<td>-1.189***</td>
<td>1.343***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-1.043**</td>
<td>0.584</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2346***</td>
<td>0.0509***</td>
<td>0.2942***</td>
<td>0.0913***</td>
<td>0.4550***</td>
<td>-0.1874***</td>
<td>-0.7841**</td>
<td>0.0604</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.0604</td>
<td>0.0604</td>
<td>-1.043***</td>
<td>-1.043***</td>
<td>1.343***</td>
<td>1.343***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-1.043**</td>
<td>0.584</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2333***</td>
<td>0.0501***</td>
<td>0.2903***</td>
<td>0.0883***</td>
<td>0.4460***</td>
<td>-0.1898***</td>
<td>-0.8852***</td>
<td>0.1464</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.1464</td>
<td>0.1464</td>
<td>-1.043***</td>
<td>-1.043***</td>
<td>1.343***</td>
<td>1.343***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-1.043**</td>
<td>0.584</td>
</tr>
<tr>
<td>ivol</td>
<td>0.2750***</td>
<td>0.0229</td>
<td>0.2865***</td>
<td>0.1022***</td>
<td>0.4031***</td>
<td>-0.1791***</td>
<td>-0.6783**</td>
<td>0.1145</td>
<td>0.0421</td>
<td>-0.1909</td>
<td>0.1145</td>
<td>0.1145</td>
<td>-1.043***</td>
<td>-1.043***</td>
<td>1.343***</td>
<td>1.343***</td>
<td>0.1647</td>
<td>0.6253***</td>
<td>0.6718**</td>
<td>0.6648***</td>
<td>-1.043**</td>
<td>0.584</td>
</tr>
</tbody>
</table>

Interacted measure $ENTSENT\_NEG$ remains highly significant
Our measure of unusualness forecasts future volatility

The best performing measure is ENTSENT_NEG, second best is negative sentiment

Much of the relevant information for future implied volatility is already impounded into options prices

Though our measure ENTSENT_NEG remains significant when we control for current vol
Theories of micro- vs macro-attention

- Allocation of attention
  - Peng and Xiong (2006)
    - Allocate attention to macro signals
  - Maćkowiak and Wiederholt (2009) [building on Sims (2003)]
    - Allocate attention to micro signals
  - Both sets of results are highly sensitive to assumptions about information structure
    - The former assumes macro information represents a larger portion of uncertainty; the latter assumes the opposite
  - Samuelson’s dictum
    - Jung and Shiller (2005)
Aggregate responses

N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Aggregate responses

Event study of aggregate variables

**top quintile:**
- VIX

**bottom quintile:**
- VIX

**top quintile:**
- ENTSENT_NEG

**bottom quintile:**
- SENTNEG
We estimate a VAR(2) model for the following system:

\[
\xi_t = \begin{bmatrix}
VIX \\
SPX_RVOL \\
ENTSENT_NEG \\
SENTNEG \\
ENTSENT_POS \\
SENTPOS \\
\end{bmatrix}
\]

Estimate structural model

\[
\xi_t = c + B_1 \xi_{t-1} + B_2 \xi_{t-2} + B_0 \mathbf{u}_t
\]

where $B_0$ is lower diagonal matrix and $\mathbf{u}_t$ are serially uncorrelated innovations with variance $I_6$
Aggregate responses summary

- Shock to unusualness interacted with negative sentiment (ENTSENT_NEG) positively affects both implied and realized vol over the ensuing 10 months, with effect peaking 4 months in the future.

- Shock to negative sentiment (SENTNEG) has no effect on implied or realized vol.

- If we change the order of the variables (SENTNEG, then ENTSENT_NEG), both shocks lead to a significant impulse response in implied and realized volatility.
  - Suggests that ENTSENT_NEG is the richer information source.

- ENTSENT_POS results are similar, but opposite sign.

- While news measures decay monotonically, the vol response is hump shaped.
Aggregate responses

Impulse response to ENTSENT_NEG with original ordering

Shock to **ENTSENT_NEG**

Shock to **SENTNEG**
Aggregate responses

Impulse response with SENTNEG before ENTSENT_NEG

Shock to SENTNEG

Shock to ENTSENT_NEG

IRF CONST 2 for SENT_NEG

IRF CONST 2 for SENT_ENT_NEG
Aggregate responses

Impulse response to ENTSENT_POS with original ordering

Shock to ENTSENT_POS

Shock to SENTPOS

IRF CONST for SENT_ENF_POS

IRF CONST for SENT_POS
Rational inattention

N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Rational inattention

Rational inattention in 1 slide

- State variable (e.g. negative log consumption) follows AR(1)

\[ X_{t+1} = \rho X_t + a u_{t+1} \]

- Can solve for a process \( Y_t \) that best tracks \( X_t \) in the sense of minimizing

\[ \mathbb{E}[(X_t - Y_t)^2] \]  \hspace{1cm} (2)

subject to an information flow constraint

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Interpretation 1</th>
<th>Interpretation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \leftrightarrow ENTSENT_{NEG} )</td>
<td>Ex post we know the largest financial firms and aggregate news flow about these sheds light on what was happening in the world over last 20 years.</td>
<td>Market participants did not have the ability to summarize news using our approach.</td>
</tr>
<tr>
<td>( Y \leftrightarrow VIX )</td>
<td>VIX reflects people’s perception of the world, which is formed subject to an information flow constraint.</td>
<td>The VIX is an optimality designed security whose payoff solves (2).</td>
</tr>
</tbody>
</table>
Rational inattention

Impulse response with unconstrained news flow

State and observation evolution for RI IRF for kappa 1.5 and rho 0.85

- State variable
- Observation

Months

- State

- Observation

kappa = 1.5
rho = 0.85
Rational inattention

Impulse response with constrained news flow

State and observation evolution for RI IRF for kappa 0.25 and rho 0.85

State variable
Observation

kappa = 0.25
rho = 0.85

Months
State

State
Observation

2 4 6 8
0.4 0.6 0.8 1.0

kappa = 0.25
rho = 0.85
State variable
Observation
Let $Y_t^*$ be the forecast of $X_t$ given information through time $t$

$$Y_t^* = \mathbb{E}[X_t|S^t]$$

The updating equations for $Y_t^*$ are

$$S_t = X_t + \sigma_\varepsilon \epsilon_t$$

$$Y_t^* = \mathbb{E}[X_t|S^{t-1}] + \Gamma (S_t - \mathbb{E}[X_t|S^{t-1}])$$

$$\mathbb{E}[X_{t+1}|S^t] = \rho Y_t^*$$

where $\Gamma$ and $\sigma_\varepsilon$ depend on the precision of the signal

Signal precision increases with the capacity of the channel
Conclusion

N-grams

Empirical implementation

Single-name responses

Aggregate responses

Rational inattention

Conclusion
Main results

- Sentiment and unusualness both matter
- Our measure of unusualness predicts volatility at the single-name and aggregate levels
- Evidence for micro-efficiency and macro-inefficiency
- Forecasting horizon is in months
- Lots more to be done!
Thank you!