Distillation of News Flow into Analysis of Stock Reactions

Junni Zhang
Cathy Chen
Wolfgang Karl Härdle
Elisabeth Bommes

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
News moves Markets...

Figure 1: Investment in: S&P 500, Sentiment Strategy

Distillation of News Flow into Analysis of Stock Reactions
Motivation

... but there is a lot of News

Distillation of News Flow into Analysis of Stock Reactions
Sentiment Projection

Many texts are numerisized via lexical projection

Goal: Accurate values for positive and negative sentiment

Figure 2: Example of Text Numerisization

Distillation of News Flow into Analysis of Stock Reactions
Sentiment Lexica

- **Opinion Lexicon** (BL)
  Hu and Liu (2004)

- **Financial Sentiment Dictionary** (LM)
  Loughran and McDonald (2011)

- **Multi-Perspective Question Answering Subjectivity Lexicon** (MPQA)
  Wilson et al. (2005)
Research Questions

- How well does numerisized sentiment explain stock reaction indicators?
- Does the lexicon matter?
Research Questions ctd

- Are there differences regarding
  1. stock reaction indicators: volatility, trading volume, returns?
  2. degree of asymmetric response (leverage effect)?
  3. high and low attention companies?
  4. specific sectors?
Outline

1. Motivation ✓
2. Data Gathering & Processing
3. Sentiment Projection
4. Panel Regression
5. Simulation
6. Conclusion
How to gather sentiment variables?

Figure 3: Flowchart of Data Gathering Process
Distillation of News Flow into Analysis of Stock Reactions
Data Gathering & Processing

NASDAQ Articles

- Web scraper for gathering text data
- Terms of Service permit web scraping
- 116,691 articles in total
- 43,459 articles about 100 selected S&P 500 stocks in 9 major GICS sectors
- Time frame: October 2009 - October 2014

Distillation of News Flow into Analysis of Stock Reactions
Sentiment Variables

- $I_{i,t}$ - article indicator (for stock $i$ on day $t$)
- $Pos_{i,t}$ - average proportion of positive words
- $Neg_{i,t}$ - average proportion of negative words
Comparison of Lexical Projections

- Average sentiment values are smaller for LM than for BL and MPQA
- Polarity: relative dominance between positive and negative sentiment

<table>
<thead>
<tr>
<th>Variable</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pos_{i,t}$ (BL)</td>
<td>88.04%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (BL)</td>
<td>10.51%</td>
</tr>
<tr>
<td>$Pos_{i,t}$ (LM)</td>
<td>55.70%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (LM)</td>
<td>40.17%</td>
</tr>
<tr>
<td>$Pos_{i,t}$ (MPQA)</td>
<td>96.26%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (MPQA)</td>
<td>2.87%</td>
</tr>
</tbody>
</table>
Correlation – Positive sentiment

Figure 4: Monthly correlation between positive sentiment: BL and LM, BL and MPQA, LM and MPQA

Distillation of News Flow into Analysis of Stock Reactions
Correlation – Negative sentiment

Figure 5: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA

Distillation of News Flow into Analysis of Stock Reactions
Comparison of Lexical Projections ctd

- BL and MPQA relatively similar
- LM only contains finance specific words
- BL and MPQA also contain more general words (e.g. "cancer")
- Combination of projections might improve results
  - PCA on sentiment scores
  - Use first principal component of $Pos_{i,t}$ and $Neg_{i,t}$
How good are the Projections?

- Random selection of 100 articles, manual labeling and comparison with lexical projections
- $BL$ and $MPQA$ underestimate negative sentiment but good in detection of positive sentiment
- $LM$ accurately estimates negative sentiment, underestimates positive sentiment
Stock Reaction Indicators


\[ \sigma_{i,t} = 0.511(u - d)^2 - 0.019 \{ c(u + d) - 2ud \} - 0.838c^2 \]  
\[ \text{with } u = \log(P_{i,t}^H) - \log(P_{i,t}^L), \quad d = \log(P_{i,t}^L) - \log(P_{i,t}^O), \]
\[ c = \log(P_{i,t}^C) - \log(P_{i,t}^O) \]

for company \( i \) on day \( t \) with \( P_{i,t}^H, P_{i,t}^L, P_{i,t}^O, P_{i,t}^C \) as highest, lowest, opening and closing stock prices, respectively.
Detrended log trading volume Girard and Biswas (2007)

\[ V_{i,t} = V_{i,t}^* - (\alpha + \beta_1 t + \beta_2 t^2) \]  

(2)

with raw log trading volume \( V_{i,t}^* \) and detrended log trading volume \( V_{i,t} \)

Returns

\[ R_{i,t} = \log(P_{i,t}^C) - \log(P_{i,t-1}^C) \]  

(3)
Panel Regression

\[
\sigma_{i,t+1} = \alpha_i + \beta_1 l_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^T X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (4)
\]

\[
V_{i,t+1} = \alpha_i + \beta_1 l_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^T X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (5)
\]

\[
R_{i,t+1} = \alpha_i + \beta_1 l_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^T X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (6)
\]

for stock \( i \) on day \( t \) where (4) to (6) are separately estimated.

\( X_{i,t} \) - control variables, \( \gamma_i \) - company specific fixed effect
Control Variables

- $R_{M,t} - \text{S&P 500 index return}$
- $VIX_t - \text{CBOE VIX}$
- $\sigma_{i,t} - \text{Range-based volatility}$
- $V_{i,t} - \text{Detrended trading volume}$
- $R_{i,t} - \text{Return}$
## Entire Panel Regression Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{i,t}$</td>
<td>−0.000</td>
<td>−0.000</td>
<td>−0.000</td>
<td>−0.000</td>
</tr>
<tr>
<td>$Pos_{i,t}$</td>
<td>−0.002</td>
<td>−0.001</td>
<td>−0.001</td>
<td>−0.001</td>
</tr>
<tr>
<td>$Neg_{i,t}$</td>
<td>0.005*</td>
<td>0.006**</td>
<td>0.004</td>
<td>0.004**</td>
</tr>
</tbody>
</table>

**Panel A: Future Volatility $\sigma_{i,t+1}$**

<table>
<thead>
<tr>
<th>Variable</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{i,t}$</td>
<td>0.047***</td>
<td>0.032***</td>
<td>0.050***</td>
<td>0.049***</td>
</tr>
<tr>
<td>$Pos_{i,t}$</td>
<td>−0.671***</td>
<td>−0.233</td>
<td>−0.618***</td>
<td>−0.470***</td>
</tr>
<tr>
<td>$Neg_{i,t}$</td>
<td>0.888***</td>
<td>0.768***</td>
<td>0.907***</td>
<td>0.589***</td>
</tr>
</tbody>
</table>

**Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$**

<table>
<thead>
<tr>
<th>Variable</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{i,t}$</td>
<td>−0.001**</td>
<td>−0.000</td>
<td>−0.000</td>
<td>−0.001**</td>
</tr>
<tr>
<td>$Pos_{i,t}$</td>
<td>0.021***</td>
<td>0.016***</td>
<td>0.016**</td>
<td>0.015***</td>
</tr>
<tr>
<td>$Neg_{i,t}$</td>
<td>−0.000</td>
<td>−0.006</td>
<td>−0.006</td>
<td>−0.003</td>
</tr>
</tbody>
</table>

**Panel C: Future Returns $R_{i,t+1}$**

*** $p$ value < 0.01, ** $0.05 < p$ value $\leq 0.01$, * $0.1 < p$ value $\leq 0.05$
Does Attention matter?

- Number of days with articles differs between firms
- Stock prices of high attention firms might incorporate news faster

\[
\text{attention ratio} \overset{\text{def}}{=} \frac{N_i}{T}
\]  \hspace{1cm} (7)

with \(N_i\) as number of days with at least one article for company \(i\) and \(T\) as total number of trading days
Grouping

Use attention ratio quartiles to group firms:

- Low: attention ratio $< Q1$
- Median: $Q1 \leq$ attention ratio $< Q2$
- High: $Q2 \leq$ attention ratio $< Q3$
- Extremely High: $Q3 \leq$ attention ratio

with $Q1$, $Q2$, $Q3$ as first, second and third quartile
## Attention Analysis Regression Results

<table>
<thead>
<tr>
<th>Attention</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Extr. High</td>
<td>Low</td>
</tr>
<tr>
<td>$I_{i,t}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Pos$_{i,t}$</td>
<td>-0.000</td>
<td>-0.001</td>
<td>-0.002</td>
</tr>
<tr>
<td>Neg$_{i,t}$</td>
<td>0.001</td>
<td>0.005***</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Panel A: Future Volatility $\sigma_{i,t+1}$**

<table>
<thead>
<tr>
<th>Attention</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Extr. High</td>
<td>Low</td>
</tr>
<tr>
<td>$I_{i,t}$</td>
<td>0.072***</td>
<td>0.033***</td>
<td>0.048***</td>
</tr>
<tr>
<td>Pos$_{i,t}$</td>
<td>-1.185***</td>
<td>-0.242</td>
<td>-1.077*</td>
</tr>
<tr>
<td>Neg$_{i,t}$</td>
<td>0.328</td>
<td>0.764**</td>
<td>0.200</td>
</tr>
</tbody>
</table>

**Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$**

<table>
<thead>
<tr>
<th>Attention</th>
<th>BL</th>
<th>LM</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Extr. High</td>
<td>Low</td>
</tr>
<tr>
<td>$I_{i,t}$</td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td>Pos$_{i,t}$</td>
<td>0.010</td>
<td>0.014</td>
<td>0.030</td>
</tr>
<tr>
<td>Neg$_{i,t}$</td>
<td>0.020</td>
<td>0.005</td>
<td>0.009</td>
</tr>
</tbody>
</table>

*** $p$ value $< 0.01$, ** $0.05 < p$ value $\leq 0.01$, * $0.1 < p$ value $\leq 0.05$
Attention Analysis Regression Results ctd

- Similar results for median and high attention groups regarding $\sigma_{i,t+1}$ and $V_{i,t+1}$
- Differences for $R_{i,t+1}$:

<table>
<thead>
<tr>
<th>Attention</th>
<th>BL Median</th>
<th>BL High</th>
<th>LM Median</th>
<th>LM High</th>
<th>MPQA Median</th>
<th>MPQA High</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{i,t}$</td>
<td>$-0.001$</td>
<td>$-0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.001^*$</td>
<td>$-0.000$</td>
</tr>
<tr>
<td>$Pos_{i,t}$</td>
<td>$0.025$</td>
<td>$0.025^*$</td>
<td>$0.032$</td>
<td>$0.034$</td>
<td>$0.039^{**}$</td>
<td>$0.026^{**}$</td>
</tr>
<tr>
<td>$Neg_{i,t}$</td>
<td>$0.008$</td>
<td>$-0.031^*$</td>
<td>$-0.037$</td>
<td>$-0.050^{***}$</td>
<td>$0.002$</td>
<td>$-0.042^{**}$</td>
</tr>
</tbody>
</table>

* *** $p$ value $< 0.01$, ** $0.05 < p$ value $\leq 0.01$, * $0.1 < p$ value $\leq 0.05$
Sector Analysis

- Compare financials sector with health care sector
- Attention ratio is high for financials sector (0.413) and low for health care sector (0.287)
- \( BL, \ MPQA \): no leverage effect of negative news for health care sector
- \( LM \): very effective in financials sector not so much in health care sector
Simulation Setup

- Evaluate the asymmetric reaction of volatility to sentiment
- $l_{i,t} \sim B(1, p_i)$
- $Pos_{i,t} \sim U(0, m_{Pos,i}), m_{Pos,i} = \max(Pos_i)$
- $Neg_{i,t} \sim U(0, m_{Neg,i}), m_{Neg,i} = \max(Neg_i)$
- Cholesky decomposition to account for correlation of $Pos_{i,t}$ and $Neg_{i,t}$
Simulation Setup ctd

- $R_{M,t} \sim G_\gamma(\mu, \sigma)$
  - Generalized Extreme Value Distribution
  - Estimate parameters from sample period
  - $\mu = 0.64$, $\sigma = 0.35$ and $\gamma = 0.20$
Simulation Setup ctd

\[ R_{i,t} - R_{f,t} = \beta_i (R_{M,t} - R_{f,t}) \]

- CAPM by Sharpe (1964) and Lintner (1965)
- Systematic risk \( \beta_i \)
- Risk-free rate \( R_{f,t} = 1\% \) p.a.
Entire Panel Results

Figure 6: Volatility Simulation for Entire Panel: Mean curve, 95% Uniform Confidence Bands

Distillation of News Flow into Analysis of Stock Reactions
Entire Panel Results ctd

- LM and MPQA: Curve for $Neg_{i,t}$ significantly differs from curve for $Pos_{i,t}$
  - Range $LM$: 0.042 - 0.094
  - Range $MPQA$: 0.051 - 0.091

- Not the case for BL
Low Attention Results

Figure 7: Volatility Simulation for Low Attention Group: Mean curve, 95% Uniform Confidence Bands

Distillation of News Flow into Analysis of Stock Reactions
Low Attention Results ctd

- Curves for $Neg_{i,t}$ do not significantly differ from curves for $Pos_{i,t}$
Extremely High Attention Results

Figure 8: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Confidence Bands
Distillation of News Flow into Analysis of Stock Reactions
Extremely High Attention Results ctd

- BL and LM: Curve for \( Neg_{t, i} \) significantly differs from curve for \( Pos_{t, i} \)
- Not the case for MPQA
Are the Bands to narrow?

- Confidence bands are based on asymptotic properties of normal distribution
- Alternative: Bootstrap confidence bands for M-Smoother by Härdle (2015)

Figure 9: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Bootstrap Confidence Bands
Distillation of News Flow into Analysis of Stock Reactions
Conclusion

- Sentiment measures: incremental information for future stock reactions
- Asymmetric impact of positive and negative sentiment
- Degree of incremental information and asymmetry is sector and attention specific
- Choice of lexicon matters
Distillation of News Flow into Analysis of Stock Reactions

Junni Zhang
Cathy Chen
Wolfgang Karl Härdle
Elisabeth Bommes

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
## Distribution over GICS sectors

<table>
<thead>
<tr>
<th>GICS Sector</th>
<th>No. Stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Discretionary</td>
<td>21</td>
</tr>
<tr>
<td>Consumer Staples</td>
<td>9</td>
</tr>
<tr>
<td>Energy</td>
<td>6</td>
</tr>
<tr>
<td>Financials</td>
<td>12</td>
</tr>
<tr>
<td>Health Care</td>
<td>15</td>
</tr>
<tr>
<td>Industrials</td>
<td>10</td>
</tr>
<tr>
<td>Information Technology</td>
<td>21</td>
</tr>
<tr>
<td>Materials</td>
<td>4</td>
</tr>
<tr>
<td>Telecommunication Services</td>
<td>2</td>
</tr>
</tbody>
</table>

Distillation of News Flow into Analysis of Stock Reactions
## Comparison of Lexical Projections

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}$</th>
<th>Max</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pos_{i,t}$ (BL)</td>
<td>0.033</td>
<td>0.012</td>
<td>0.134</td>
<td>0.025</td>
<td>0.032</td>
<td>0.040</td>
<td>88.04%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (BL)</td>
<td>0.015</td>
<td>0.010</td>
<td>0.091</td>
<td>0.008</td>
<td>0.014</td>
<td>0.020</td>
<td>10.51%</td>
</tr>
<tr>
<td>$Pos_{i,t}$ (LM)</td>
<td>0.014</td>
<td>0.007</td>
<td>0.074</td>
<td>0.009</td>
<td>0.013</td>
<td>0.018</td>
<td>55.70%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (LM)</td>
<td>0.012</td>
<td>0.009</td>
<td>0.085</td>
<td>0.006</td>
<td>0.011</td>
<td>0.016</td>
<td>40.17%</td>
</tr>
<tr>
<td>$Pos_{i,t}$ (MPQA)</td>
<td>0.038</td>
<td>0.012</td>
<td>0.134</td>
<td>0.031</td>
<td>0.038</td>
<td>0.045</td>
<td>96.26%</td>
</tr>
<tr>
<td>$Neg_{i,t}$ (MPQA)</td>
<td>0.013</td>
<td>0.008</td>
<td>0.133</td>
<td>0.007</td>
<td>0.012</td>
<td>0.017</td>
<td>2.87%</td>
</tr>
</tbody>
</table>

Sample mean, sample standard deviation, maximum value, 1st, 2nd and 3rd quartiles, and polarity as relative dominance between positive and negative sentiment.
# Classification Evaluation

<table>
<thead>
<tr>
<th>Manual Label</th>
<th>BL Label</th>
<th>LM Label</th>
<th>MPQA Label</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pos</td>
<td>Neg</td>
<td>Neu</td>
</tr>
<tr>
<td>Pos</td>
<td>56</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Neg</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Neu</td>
<td>22</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>
Algorithm: Bootstrap Confidence Bands I

1) Compute $\hat{m}_h(x)$ by using the curve estimator proposed by Nadaraya (1964) and Watson (1964):

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}$$

where $K_h(u) = \varphi(u/h)/h$ denotes the Gaussian Kernel and set $\hat{\varepsilon}_i \stackrel{def}{=} Y_i - \hat{m}_h(X_i)$. To ensure robustness against outliers, this estimator is adjusted as proposed in Brillinger (1977).
Algorithm: Bootstrap Confidence Bands II

2) Compute the estimated conditional distribution function
   \( \hat{F}_{(\varepsilon|X)}(\cdot) \) with Gaussian kernel.

3) Construct \( j = 1, \ldots, J \) samples by generating the random variables
   \( \varepsilon_i^* \sim \hat{F}_{(\varepsilon|X=X_i)} \) with \( i = 1, \ldots, n \) for each sample.
   Compute
   \[
   Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*
   \]
   with \( g \) chosen such that \( \hat{m}_g(X_i) \) is slightly oversmoothed.
Algorithm: Bootstrap Confidence Bands

4) For each bootstrap sample \( \{X_i, Y_i^*\}_{i=1}^n \), compute \( \hat{m}_{h,g}(\cdot) \) and the random variable

\[
d_j \overset{\text{def}}{=} \sup_{x \in B} \left| \hat{m}_{h,g}(x) - \hat{m}_g(x) \right| \sqrt{\hat{f}_X(x) \hat{f}_{(\varepsilon|X)}(x)} / \sqrt{\hat{E}_{(\varepsilon|X)}\{\psi^2(\varepsilon)\}},
\]

for a finite number of points in the compact set \( B \). Both \( \hat{f}_{(\varepsilon|X)}(x) \) and \( \hat{E}_{(\varepsilon|X)}\{\psi^2(\varepsilon)\} \) are computed using the estimated residuals \( \hat{\varepsilon}_i \). \( \psi(\cdot) \) denotes the \( \psi \)-function by Huber(2011) with \( \psi(u) = \max\{-c, \min(u, c)\} \) for \( c > 0 \).
Algorithm: Bootstrap Confidence Bands IV

5) Calculate the $1 - \alpha$ quantile $d^*_\alpha$ of $d_1, \ldots, d_J$.

6) Construct the bootstrap uniform band centered around $\hat{m}_h(x)$

$$\hat{m}_h(x) \pm \left[ \sqrt{\hat{f}_X(x)\hat{f}_{\varepsilon|X}(x)} / \sqrt{\hat{E}_{\varepsilon|X}\{\psi^2(\varepsilon)\}} \right]^{-1} d^*_\alpha.$$

Distillation of News Flow into Analysis of Stock Reactions
For Further Reading

- Tobias Oetiker, Hubert Partl, Irene Hyna and Elisabeth Schlegl
  *The Not So Short Introduction to \LaTeX\!*2e*
  available on [www.ctan.org](http://www.ctan.org), 2008

- Scott Pakin
  *The Comprehensive \LaTeX\!Symbol List*
  available on [www.ctan.org](http://www.ctan.org), 2008

- Frank Mittelbach and Michel Goossens
  *The \LaTeX! Companion – 2nd ed.*
  Addison-Wesley, 2004
For Further Reading

- **Mark Trettin and Jürgen Fenn**
  *An essential guide to \LaTeX\usage*
  available on www.ctan.org, 2007

- **Wikipedia Wiki Books**
  *LaTeX-Wörterbuch: InDeX*
  available on www.wikipedia.de

- **Till Tantau**
  *User Guide to the Beamer Class, Version 3.07*
  available on www.sourceforge.net, 2007