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...

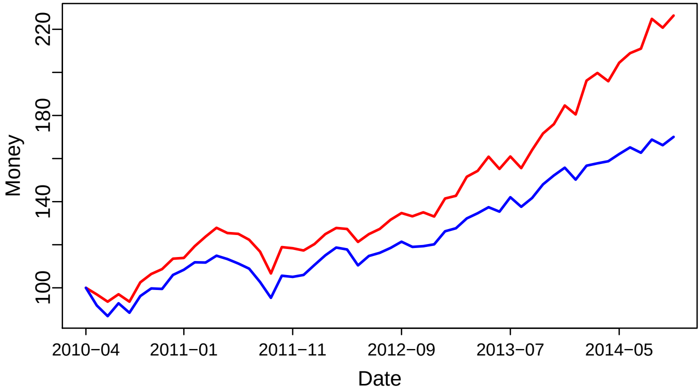

Figure 1: Investment in: S&P 500, Sentiment Strategy

... but there is a lot of News

Sentiment Projection

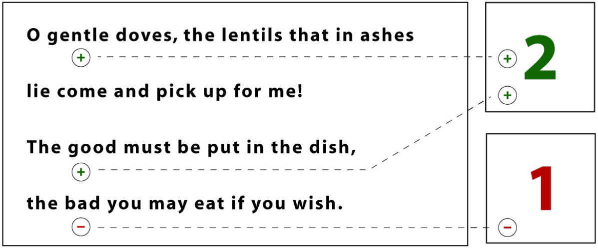

Figure 2: Example of Text Numerisization

- Many texts are numerisized via lexical projection
- Goal: Accurate values for positive and negative sentiment

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?

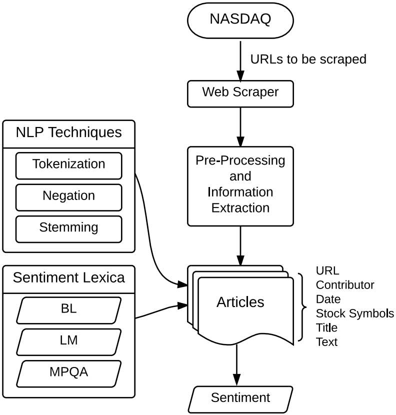

Figure 3: Flowchart of Data Gathering Process

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 GICS distribution
- Time frame: October 2009 - October 2014

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

Variable	Polarity
$Pos_{i,t}$ (BL)	88.04%
$Neg_{i,t}$ (BL)	10.51%
$Pos_{i,t}$ (LM)	55.70%
$Neg_{i,t}$ (LM)	40.17%
$Pos_{i,t}$ (MPQA)	96.26%
$Neg_{i,t}$ (MPQA)	2.87%

Summary Statistics

Correlation - Positive sentiment

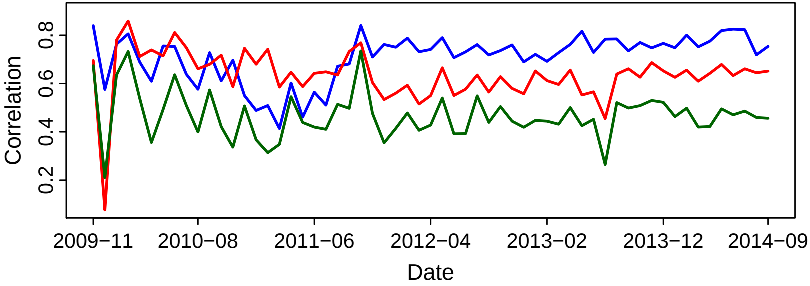

Figure 4: Monthly correlation between positive sentiment: BL and LM, BL and MPQA, LM and MPQA

Correlation - Negative sentiment

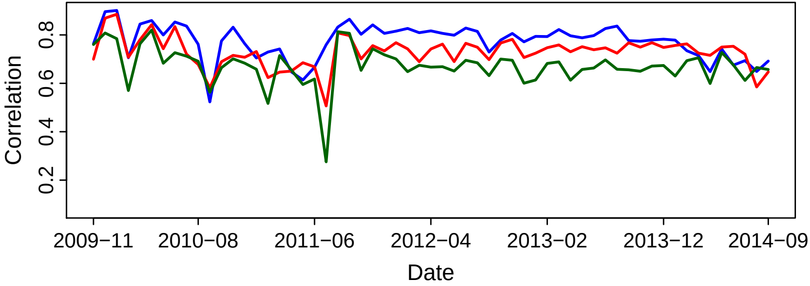

Figure 5: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA

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

Classification Evaluation Table

Stock Reaction Indicators

Range-based measure of volatility by Garman and Klass (1980)

$$\sigma_{i,t} = 0.511(u-d)^2 - 0.019\left\{c(u+d) - 2ud\right\} - 0.838c^2 \quad (1)$$

$$\text{with } u = \log(P^H_{i,t}) - \log(P^L_{i,t}), \quad d = \log(P^L_{i,t}) - \log(P^O_{i,t}),$$
$$c = \log(P^C_{i,t}) - \log(P^O_{i,t})$$

for company i on day t with $P^H_{i,t}$, $P^L_{i,t}$, $P^O_{i,t}$, $P^C_{i,t}$ as highest, lowest, opening and closing stock prices, respectively.

Detrended log trading volume Girard and Biswas (2007)

$$V_{i,t} = V^*_{i,t} - (\alpha + \beta_1 t + \beta_2 t^2) \quad (2)$$

with raw log trading volume $V^*_{i,t}$ and detrended log trading volume $V_{i,t}$

Returns

$$R_{i,t} = \log(P^C_{i,t}) - \log(P^C_{i,t-1}) \quad (3)$$

Panel Regression

$$
\begin{aligned}
\sigma_{i,t+1} &= \alpha_i + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} && (4) \\
V_{i,t+1} &= \alpha_i + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} && (5) \\
R_{i,t+1} &= \alpha_i + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} && (6)
\end{aligned}
$$

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

$X_{i,t}$ - control variables, γ_i - company specific fixed effect

Control Variables

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

Entire Panel Regression Results

Variable	BL	LM	MPQA	PCA
	Panel A: Future Volatility $\sigma_{i,t+1}$			
$I_{i,t}$	-0.000	-0.000	-0.000	-0.000
$Pos_{i,t}$	-0.002	-0.001	-0.001	-0.001
$Neg_{i,t}$	0.005^{*}	0.006^{**}	0.004	0.004^{**}
	Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$			
$I_{i,t}$	0.047^{***}	0.032^{***}	0.050^{***}	0.049^{***}
$Pos_{i,t}$	-0.671^{***}	-0.233	-0.618^{***}	-0.470^{***}
$Neg_{i,t}$	0.888^{***}	0.768^{***}	0.907^{***}	0.589^{***}
	Panel C: Future Returns $R_{i,t+1}$			
$I_{i,t}$	-0.001^{**}	-0.000	-0.000	-0.001^{**}
$Pos_{i,t}$	0.021^{***}	0.016^{***}	0.016^{**}	0.015^{***}
$Neg_{i,t}$	-0.000	-0.006	-0.006	-0.003

*** p value < 0.01, ** $0.05 < p$ value ≤ 0.01, * $0.1 < p$ value ≤ 0.05

Does Attention matter?

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

$$\text{attention ratio} \stackrel{def}{=} N_i / T \tag{7}$$

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

Attention	BL Low	BL Extr. High	LM Low	LM Extr. High	MPQA Low	MPQA Extr. High
	Panel A: Future Volatility $\sigma_{i,t+1}$					
$I_{i,t}$	0.000	0.000	0.000	-0.000	0.000	0.000
$Pos_{i,t}$	-0.000	-0.001	-0.002	-0.002	-0.001	-0.001
$Neg_{i,t}$	0.001	0.005***	0.001	0.007***	0.001	0.004**
	Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$					
$I_{i,t}$	0.072***	0.033***	0.048***	0.025**	0.067***	0.049***
$Pos_{i,t}$	-1.185***	-0.242	-1.077*	0.327	-0.815**	-0.623*
$Neg_{i,t}$	0.328	0.764**	0.200	0.709**	-0.900	0.936**
	Panel C: Future Returns $R_{i,t+1}$					
$I_{i,t}$	-0.000	-0.000	-0.000	-0.001	0.000	0.000
$Pos_{i,t}$	0.010	0.014	0.030	0.030	0.010	-0.007
$Neg_{i,t}$	0.020	0.005	0.009	-0.025*	-0.011	0.007

*** p value < 0.01, ** $0.05 < p$ value ≤ 0.01, * $0.1 < p$ value ≤ 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 $R_{i,t+1}$:

	BL		LM		MPQA	
Attention	Median	High	Median	High	Median	High
	Panel C: Future Returns $R_{i,t+1}$					
$I_{i,t}$	-0.001	-0.000	0.000	0.000	0.001*	-0.000
$Pos_{i,t}$	0.025	0.025*	0.032	0.034	0.039**	0.026**
$Neg_{i,t}$	0.008	-0.031*	-0.037	-0.050***	0.002	-0.042**

*** p value < 0.01, ** $0.05 < p$ value ≤ 0.01, * $0.1 < p$ value ≤ 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
- $I_{i,t} \sim B(1, p_i)$
- $Pos_{i,t} \sim U(0, m_{Pos,i}),\ m_{Pos,i} = \max(Pos_i)$
- $Neg_{i,t} \sim U(0, m_{Neg,i}),\ m_{Neg,i} = \max(Neg_i)$
- 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 β_i
  - Risk-free rate $R_{f,t} = 1\%$ p.a.

Entire Panel Results

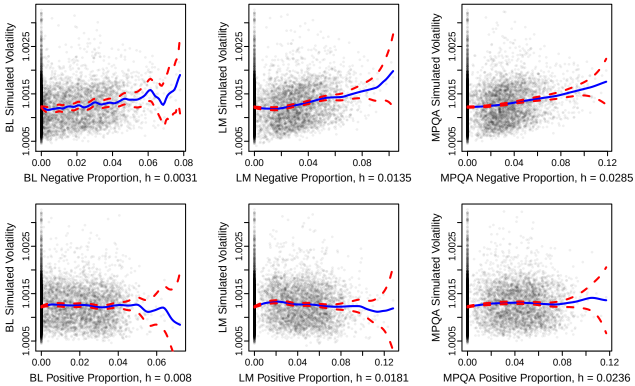

Figure 6: Volatility Simulation for Entire Panel: Mean curve, 95% Uniform Confidence Bands

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

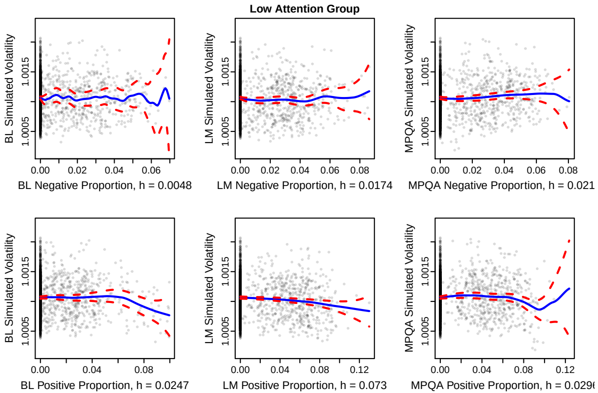

Figure 7: Volatility Simulation for Low Attention Group: Mean curve, 95% Uniform Confidence Bands

Low Attention Results ctd

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

Extremely High Attention Results

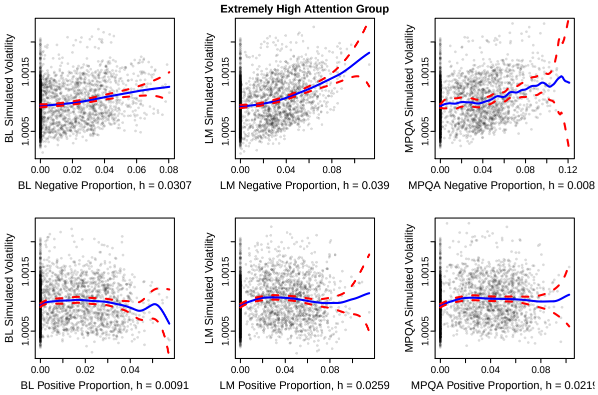

Figure 8: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Confidence Bands

Extremely High Attention Results ctd

- *BL* and *LM*: Curve for $Neg_{i,t}$ significantly differs from curve for $Pos_{i,t}$
- 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) Algorithm

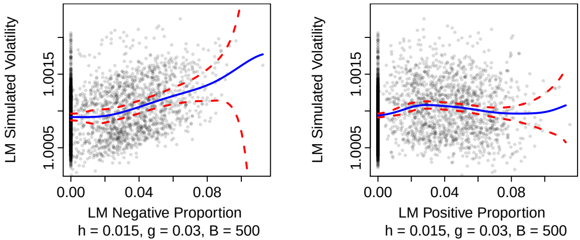

Figure 9: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Bootstrap Confidence Bands

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

GICS Sector	No. Stocks
Consumer Discretionary	21
Consumer Staples	9
Energy	6
Financials	12
Health Care	15
Industrials	10
Information Technology	21
Materials	4
Telecommunication Services	2

Back

Comparison of Lexical Projections

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

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

Back

Classification Evaluation

Manual Label	BL Label			LM Label			MPQA Label			
	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Total
Pos	56	4	1	41	12	8	61	0	0	61
Neg	9	2	1	0	9	3	9	2	1	12
Neu	22	5	0	10	15	2	26	0	1	27
Total	87	11	2	51	36	13	96	2	2	100

Back

Algorithm: Bootstrap Confidence Bands

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 III

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

$$d_j \stackrel{def}{=} \sup_{x \in B} [|\hat{m}^*_{h,g}(x) - \hat{m}_g(x)| \sqrt{\hat{f}_X(x) \hat{f}_{(\varepsilon|X)}(x)\}} / \sqrt{\widehat{\mathsf{E}}_{\varepsilon|X}\{\psi^2(\varepsilon)\}}],$$

$$j = 1, \dots, J$$

for a finite number of points in the compact set B. Both $\hat{f}_{(\varepsilon|X)}(x)$ and $\widehat{\mathsf{E}}_{\varepsilon|X}\{\psi^2(\varepsilon)\}$ are computed using the estimated residuals $\hat{\varepsilon}_i$. $\psi(\cdot)$ denotes the ψ-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^*_α of $d_1, \ldots, d_J$.
6) Construct the bootstrap uniform band centered around $\hat{m}_h(x)$

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

Back

For Further Reading

Tobias Oetiker, Hubert Partl, Irene Hyna and Elisabeth Schlegl
The Not So Short Introduction to LaTeX2e
available on www.ctan.org, 2008

Scott Pakin
The Comprehensive LaTeXSymbol List
available on 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 LaTeX2e 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