Textual Sentiment, Option Information and Stock Return Predictability

Cathy Yi-Hsuan Chen
Matthias R. Fengler
Wolfgang Karl Härdle
Yanchu Liu

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
University of St. Gallen, Switzerland
Lingnan College, Sun Yat-sen University, China

http://lvb.wiwi.hu-berlin.de
http://www.mathstat.unisg.ch/
http://www.lingnan.sysu.edu.cn

Sentiment moves stock markets

- Growing evidence shows that textual sentiment provides incremental information about future stock returns.

  Confirmed at index levels as well as single-stock levels.

- Antweiler & Frank (2004), Tetlock (2007), Tetlock (2011), Hillert et al. (2014), Zhang et al. (2016), among others.

What about sentiment and options markets?

- Han (2008): aggregate sentiment proxies (*Investors Intelligence* survey, CFTC reported long-short futures, Sharpe's (2002) index valuation errors) predict risk neutral skewness of index options.
- Prediction power cannot be explained by "rational" option pricing models.

Options market and stock market

- Dennis and Mayhew (2002), Xing et al. (2010): option data characteristics (skew, implied volatility) predict stock returns

  Hypothesis:
  private information about stocks can be best exploited via the option market because it's easier to leverage and short-sell.

  Therefore options market may lead stock markets in terms of price discovery.

Given sentiment predicts both stock returns and option data, is there still room for the private information hypothesis in option markets?

Maybe it's all just a common sentiment factor that get's internalized at different speed in the different markets.

Requires a joint study of

Textual Sentiment, Option Information and Stock Return Predictability

This research

- Extend Han's (2008) ideas:
  - Study reaction of standard of single-stock options to news
  - Use language processing tools for sentiment construction
- Investigate influence of option market variables in presence of news sentiment (Xing et al.'s hypothesis)
- Study source of option markets predictability:
  Inside information? Internalized investor sentiment? Both?

Current literature

$$
\begin{tikzcd}
B_t \arrow[r] & R_{t+1} \\
OC_t \arrow[ur] & OC_{t+1}
\end{tikzcd}
$$

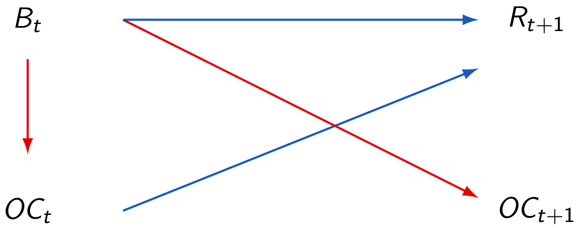
B_t is sentiment, OC_t an option market variable, R_t a stock return

This work

$$\begin{tikzcd}
B_t \arrow[d] \arrow[dr] & R_{t+1} \\
OC_t & OC_{t+1}
\end{tikzcd}$$

B_t is sentiment, OC_t an option market variable, R_t a stock return

This work



B_t is sentiment, OC_t an option market variable, R_t a stock return

Findings

- Our sentiment proxies predict single-stock option market variables
  - Both firm-specific sentiment and aggregate sentiment
  - Aggregate negative sentiment is a strong predictor
- Sentiment proxies predict single stock returns
- Asymmetry of informational relevance of news:
  - Overnight information more relevant than trading day information
  - Possibly due to a different thematic coverage and more complex topics.

Findings

- Option market variables remain relevant predictors of stock returns in presence of sentiment
  - Aggregate sentiment is a relevant factor for single stock returns
  - Option market variables where sentiment is partialled out remain significant predictors.

Outline

1. Motivation ✓
2. Data collection
3. Text analytics
4. Sentiment projection
5. Topic model
6. Panel regressions
7. Summary

Sentiment extraction from news data

There is a lot of news...

Dimensions of news

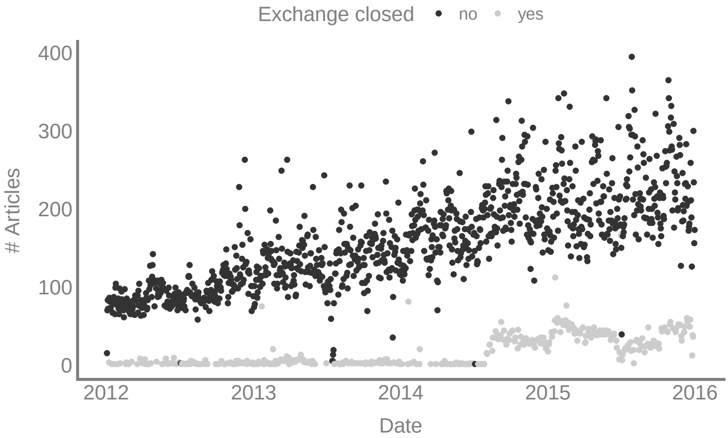
- Source of news
  - Official channels: government, federal reserve bank/central bank, financial institutions
  - Internet: blogs, social media, message boards
- Type of news
  - Scheduled vs. non-scheduled
  - Expected vs. unexpected
  - Event-specific vs. continuous news flows

Data

Sentiment variables: distilled from Nasdaq articles

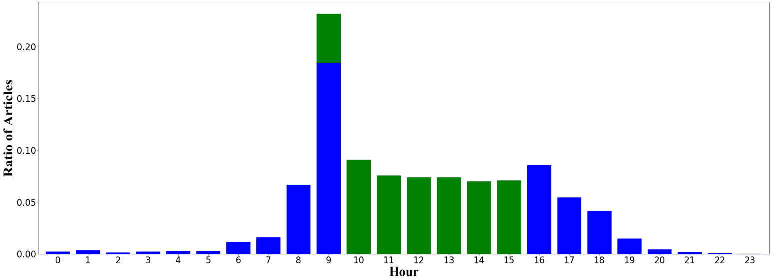
- Terms of Service permit web scraping
- Currently $>$ 580k articles between October 2009 and January 2017
- Data available at RDC
- Analysis is on data from 2012-2015

Number of articles per trading day



Black: # articles on a trading day; grey: # articles on weekend, holiday

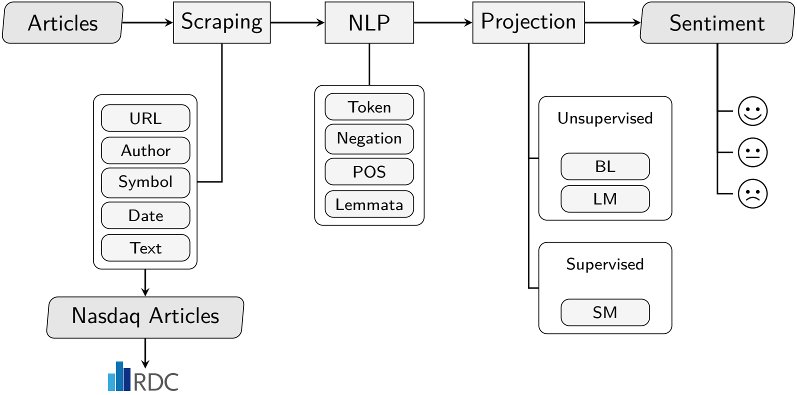
Hourly distribution



In total we process

- 119,680 articles, out of which 6,600 articles (i.e., 5.51%) are posted on non-trading days (excluded)
- Out of 113,080 articles 50.26% are posted during trading hours and 49.74% during overnights.

Extracting sentiment from text



Sentiment analysis

Strategies:

- Lexica projection : positive, neutral and negative
- Machine learning : text classification

Based on:

- *Financial Sentiment Dictionary* (LM)
  Loughran and McDonald (JF, 2011)
- *Financial Phrase Bank* (LM)
  Malo et al. (2014)

Lexicon Correlation

Unsupervised projection

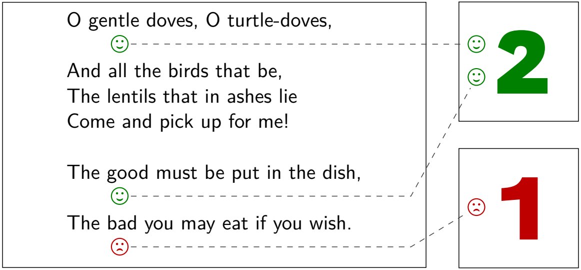


Figure: Example of Text Numerisization

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

Examples

Lexicon-based sentiment

Consider sentence i in some document, positive sentiment Pos_i, positive lexicon entries W_j ($j = 1, \ldots, J$) and count frequency of those entries w_j:

$$Pos_i = n_i^{-1} \sum_{j=1}^{J} \mathbf{I} \left(W_j \in L \right) w_j$$

with n_i: number of words in document i (e.g. sentence)

Equivalent calculation of negative sentiment Neg_i

Sentence-level polarity

For sentence i, we compute the sentence-level polarity by:

$$Pol_i = \begin{cases} 1, & \text{if } Pos_i > Neg_i \\ 0, & \text{if } Pos_i = Neg_i \\ -1, & \text{if } Pos_i < Neg_i \end{cases}.$$

Then, at the document level, we calculate,

$$FP = n^{-1} \sum_{i=1}^{n} \mathbf{I}(Pol_i = 1)$$

$$FN = n^{-1} \sum_{i=1}^{n} \mathbf{I}(Pol_i = -1),$$

where n is the number of sentences in the document.

Supervised projection

- Training data: Financial Phrase Bank of Malo et al. (2014)
  - Sentence-level annotation of financial news
  - **Manual annotation** of 5,000 sentences by 16 annotators incorporates human knowledge
  - Example: “profit” with different semantic orientations
    - Neutral in “profit was 1 million”
    - Positive in “profit increased from last year”

Regularized linear models (RLM)

- Training data $(X_1, y_1) \dots (X_n, y_n)$ with $X_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$
- Linear scoring function $s(X) = \beta^\top X$ with $\beta \in \mathbb{R}^p$

Example

Regularized training error:

$$n^{-1} \sum_{i=1}^{n} \underbrace{L\{y_i, s(X)\}}_{\text{Loss Function}} + \lambda \underbrace{R(\beta)}_{\text{Regularization Term}} \tag{1}$$

with hyperparameter $\lambda \geq 0$

RLM estimation

- Optimize via Stochastic Gradient Descent More
- 5-fold cross validation More
- Oversampling More
- Choice of: $L(\cdot), R(\cdot), \lambda$, X (n-gram range, features) ...
- Three categories: one vs. all sub-models

Model accuracy - polarity

Supervised Learning

- Chosen model: Hinge loss, L1 norm, $\lambda = 0.0001$, ...
- Mean accuracy (oversampling): 0.80
- Mean accuracy (normal sample): 0.82

Lexicon-based

- Mean accuracy BL: 0.58
- Mean accuracy LM: 0.64

So, we adopt the supervised learning methodology hereafter.

Confus. Matrix

Sentence-level and document-level polarity

After training: Each document i is split up into its sentences j and the corresponding score is calculated.

Yields a predictor for the polarity of sentence j, Pol_j:

For each document, these scores are aggregated to

$$FP = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = 1)$$

$$FN = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = -1)\,,$$

where n is the number of sentences in the document.

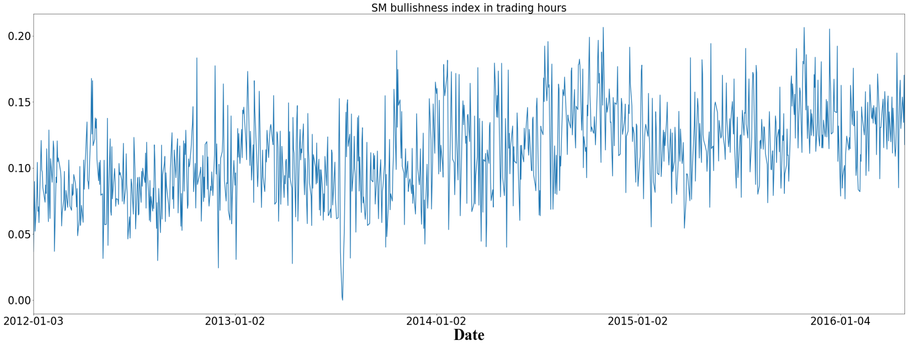
Bullishness

$$B = \log\left\{\frac{1 + n^{-1}\sum_{j=1}^{n} \mathbf{I}\left(Pol_j = 1\right)}{1 + n^{-1}\sum_{j=1}^{n} \mathbf{I}\left(Pol_j = -1\right)}\right\} \quad (2)$$

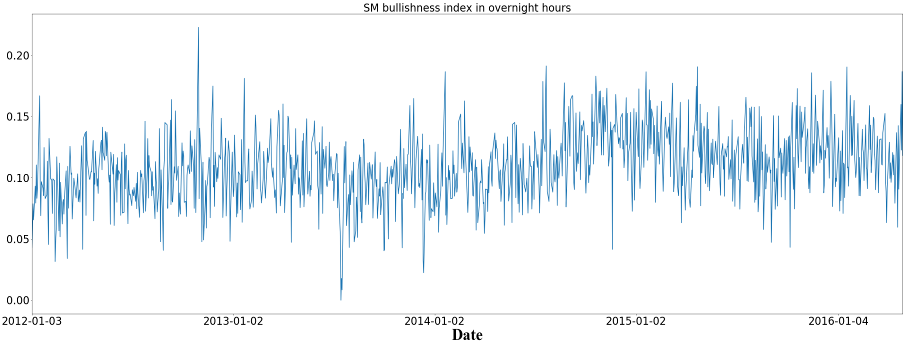
by Antweiler and Frank (JF, 2004) with $j = 1, \ldots, n$ sentences in document.

- $B_{i,t}$ accounts for bullishness of company i on day t
- Consider $BN_{i,t} = -\mathbf{I}\left(B_{i,t} < 0\right) B_{i,t}$

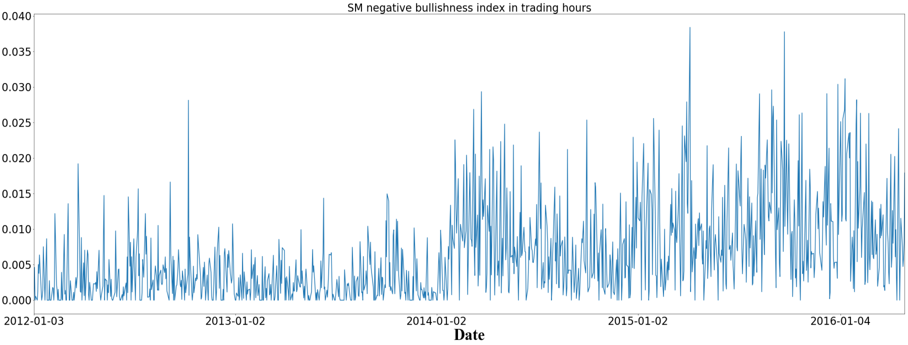
trading B_{idx}



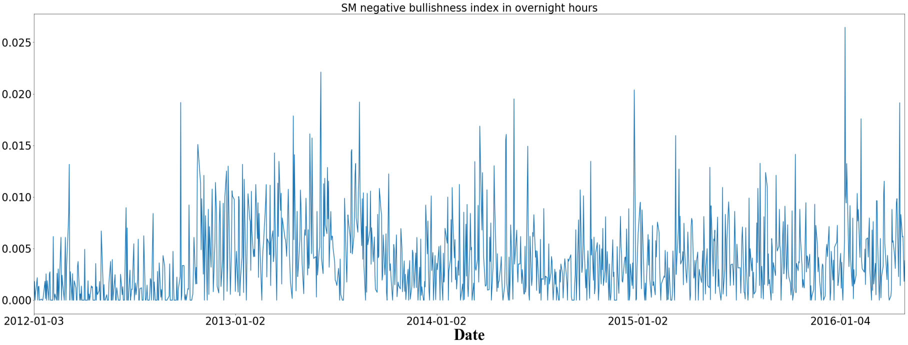
overnight B_{idx}^{on}



trading BN_{idx}



overnight BN_{idx}^{on}



How do trading-day/overnight articles differ?

- Overnight information is more informative than trading-day information. Why?
- Uncover the thematic coverage of the alternate news archives using a statistical topic model

Latent Dirichlet Allocation

LDA is a topic model suggested by Blei, Ng and Jordan (2003).

Structure:

- Documents are random mixtures over latent topics.
- A topic is a distribution over a fixed vocabulary (generated before the documents).
- A document may feature several topics.

Details LDA

LDA: overnight archive

	Topics and most frequent words						
	1	2	3	4	5	6	7
Topics	*Dividends*	*Inv. stratg.*	*Earnings*	*Equities*	*Asset mgmt*	*Econ. Outlook*	*Charts*
Top 15 words	**dividend**	**stock**	**earnings**	tale	**fund**	**stocks**	**average**
	ex	reasons	**estimates**	tape	**income**	buy	**moving**
	date	**focus**	follow	**continue**	**municipal**	**oil**	day
	scheduled	**great**	**history**	**higher**	**nuveen**	higher	**cross**
	corporation	**investors**	**indicator**	**shares**	**dividend**	**week**	**bullish**
	september	choice	reaction	**focus**	**ex**	best	notable
	june	**value**	**sensitive**	**estimates**	**scheduled**	**news**	makes
	march	**jumps**	**revenues**	march	**date**	**data**	**critical**
	november	session	beat	**surge**	high	lower	**breaks**
	august	**growth**	beats	**strong**	new	**ahead**	key
	trust	**momentum**	season	**value**	**eaton**	**watch**	**level**
	february	rises	**surprise**	**great**	**vance**	today	crosses
	december	right	revenue	**growth**	**trust**	**china**	alert
	july	adds	strong	**falls**	quality	**dividend**	**crossove**
	october	moves	misses	**holdings**	ii	**growth**	dow

LDA: overnight archive, ctd.

Topics and most frequent words							
3	4	5	6	7	8	9	10
Earnings	*Equities*	*Asset mgmt*	*Econ. Outlook*	*Charts*	*Anal. Roundup*	*Sectors*	*Market*
earnings	tale	**fund**	**stocks**	**average**	**analyst**	**update**	**market**
stimates	tape	**income**	buy	**moving**	**blog**	**sector**	**report**
follow	**continue**	**municipal**	**oil**	day	**growth**	**energy**	pre
history	**higher**	**nuveen**	higher	**cross**	new	**health**	**nasdaq**
indicator	**shares**	**dividend**	**week**	**bullish**	data	**care**	**index**
reaction	**focus**	**ex**	best	notable	**beat**	**financial**	**close**
sensitive	**estimates**	**scheduled**	**news**	makes	**shares**	**consumer**	active
revenues	march	**date**	**data**	**critical**	**energy**	ung	**composite**
beat	**surge**	high	lower	**breaks**	**high**	uso	**closes**
beats	**strong**	new	**ahead**	key	week	**technology**	**points**
season	**value**	**eaton**	**watch**	**level**	miss	**close**	qqq
surprise	**great**	**vance**	today	crosses	loss	closing	**aapl**
revenue	**growth**	**trust**	**china**	alert	**roundup**	**oil**	**bac**
strong	**falls**	quality	**dividend**	**crossover**	**revenues**	partners	xiv
misses	**holdings**	ii	**growth**	dow	estimates	dis	tvix

LDA: trading-day archive

	Topics and most frequent words						
	1	2	3	4	5	6	7
Topics	*Press rel.*	*Earnings 1*	*Funds*	*Option trades*	*Charts*	*Sectors*	*Dividend*
	analyst	**earnings**	**etf**	**options**	**average**	update	**stock**
	blog	**revenues**	detected	**trading**	**moving**	**sector**	**reminde**
	zacks	**beat**	big	**using**	day	**energy**	market
	highlights	**estimates**	**inflow**	week	cross	**financial**	**preferre**
	releases	**beats**	**inflows**	**interesting**	**bullish**	**technology**	today
	press	**miss**	**outflow**	**earn**	notable	**consumer**	**series**
	energy	season	**outflows**	commit	**critical**	**health**	news
Top 15 words	**group**	**report**	notable	**buy**	makes	**care**	**ex**
	holdings	view	large	annualized	**breaks**	mid	**cumulativ**
	international	store	noteworthy	**available**	key	**market**	**dividend**
	high	**sales**	alert	begin	**crosses**	afternoon	interestin
	american	misses	experiences	**purchase**	level	day	corp
	loss	tops	**ishares**	october	**crossover**	laggards	**roundup**
	week	**surprise**	**etfs**	january	alert	**oil**	**redeemab**
	airlines	**revenue**	**spdr**	november	option	morning	non

LDA: trading-day archive, ctd.

Topics and most frequent words							
3	4	5	6	7	8	9	10
Funds	*Option trades*	*Charts*	*Sectors*	*Dividends*	*Equities*	*Earnings 2*	*Share types*
etf	**options**	**average**	update	**stock**	**stocks**	**indicator**	**shares**
detected	**trading**	**moving**	**sector**	**reminder**	**buy**	**earnings**	cross
big	**using**	day	**energy**	market	**new**	follow	**yield**
inflow	week	cross	**financial**	**preferred**	strong	**history**	**series**
inflows	**interesting**	**bullish**	**technology**	today	**oil**	**reaction**	mark
outflow	**earn**	notable	**consumer**	**series**	mid	**sensitive**	**preferred**
outflows	commit	**critical**	**health**	news	**sell**	corp	dma
notable	**buy**	makes	**care**	**ex**	**etfs**	**corporation**	**dividend**
large	annualized	**breaks**	mid	**cumulative**	european	company	today
noteworthy	**available**	key	**market**	**dividend**	**adrs**	international	mid
alert	begin	**crosses**	afternoon	interesting	day	**group**	**cumulative**
experiences	**purchase**	level	day	corp	news	systems	**ex**
ishares	october	**crossover**	laggards	**roundup**	**market**	technology	higher
etfs	january	alert	**oil**	**redeemable**	gains	**holdings**	afternoon
spdr	november	option	morning	non	higher	technologies	reminder

Option markets' reaction to sentiment

- Fixed-effect panel regression with IV

$$OC_{it} = \alpha + \gamma_i + \beta_1 B_{it} + \beta_2^\top X_{it} + \varepsilon_{it} \quad (3)$$

- $OC_{it} \in \{Skew_{it}, IVol_{it}, OTM_{it}\}$: option characteristic
- X_{it}: the vector of control variables More Information

Endogeneity

- Sentiment for single stocks and reaction in options market could be due to a common cause.
- Need to assert that NASDAQ news/articles are the only source of news.
- Idea:
  - Use lagged $B_{i,t-1}$, $B_{idx,t-1}$, $BN_{idx,t-1}$ as instruments

OCs and sentiment in trading hours

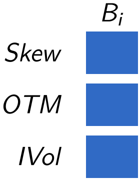


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- IV regressions with constant, fixed effects, and FF1-5 factors
- instrument: $B_{i,t-1}$
- Blue (negative sign);Red (positive sign)

OCs and sentiment in trading hours

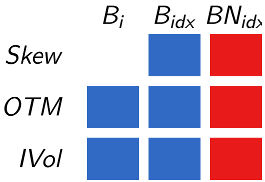


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- IV regressions with constant, fixed effects, and FF1-5 factors
- instrument: $B_{i,t-1}$, $B_{idx,t-1}$, $BN_{idx,t-1}$
- Blue (negative sign);Red (positive sign)

Option markets' reaction: summary

- Standard endogeneity tests (Durbin, Hausman-Wu) reject that B_{it} is exogenous
- *Skew*, *IVol* and *OTM* react to investor sentiment
- Higher B results in a flatter *Skew*, lower *OTM* and *IVOl*
- Higher B_{idx} results in a flatter *Skew*, lower *OTM* and *IVOl*
- Higher BN_{idx} results in a steeper *Skew*, higher *OTM* and *IVOl*

Stock return predictability: Option variables v.s. sentiment index

Pooled OLS regressions

$$R_{i,t+1} = \alpha + \beta_1 OC_{it} + \beta_2 B_{i,t} + \beta_3 B_{idx,t} + \beta_4 BN_{idx,t} + \beta_5 B^{on}_{i,t} + \beta_6 B^{on}_{idx,t} + \beta_7 BN^{on}_{idx,t} + \beta_8^\top X_{it} + \varepsilon_{it}$$

- Xing et al. (JFQA, 2010) only use OC_{it}
- Incremental predictability from sentiment index

Stock return predictability: Option variables

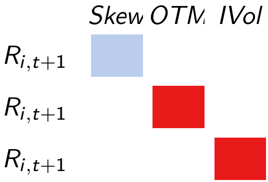


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Blue (negative sign);Red (positive sign)

Stock return predictability: Option variables and sentiment

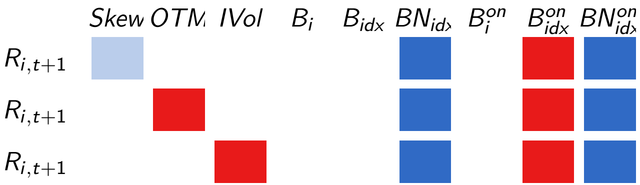


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Blue (negative sign);Red (positive sign)

Stock return predictability ctd

- Confirms Xing et al. (JFQA, 2010)'s results on the predictability of *Skew*
- Stock-specific sentiment insignificant
- Negative aggregate trading and overnight sentiment carry significant predictive content in presence of options market variables
- Aggregate overnight sentiment is a good predictor too.

Decompose option variables: Sentiment-related v.s. non-public part

Extract sentiment component from option market variables.

- Regress OC on sentiment and controls to get residuals:

$$OC_{i,t} = \alpha + \theta^\top \mathbf{B}_t + \beta^\top X_{i,t} + \epsilon^i_{OC,t}$$

- $\{Skew_{i,t}, Put_{i,t}, IV_{i,t}\} \in OC_{i,t}$.
  $\mathbf{B}_t = (B_{i,t}, B_{idx,t}, BN_{idx,t}, B^{on}_{i,t}, B^{on}_{idx,t} BN^{on}_{idx,t})^\top$.
- $\epsilon^i_{OC,t}$: residual term as a proxy for non-public information embedded in options data

Use residuals in the regression:

Pooled OLS regressions

$$\begin{aligned} R_{i,t+1} = \alpha + \beta_1 \epsilon^{i}_{OC,t} + \beta_2 B_{i,t} + \beta_3 B_{idx,t} + \beta_4 BN_{idx,t} \\ + \beta_5 B^{on}_{i,t} + \beta_6 B^{on}_{idx,t} + \beta_7 BN^{on}_{idx,t} + \beta_8^\top X_{it} + \varepsilon_{it} \end{aligned}$$

Stock return predictability: Option variables and sentiment

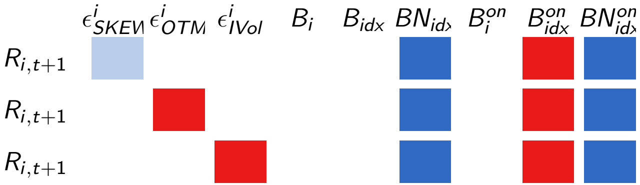


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Blue (negative sign);Red (positive sign)

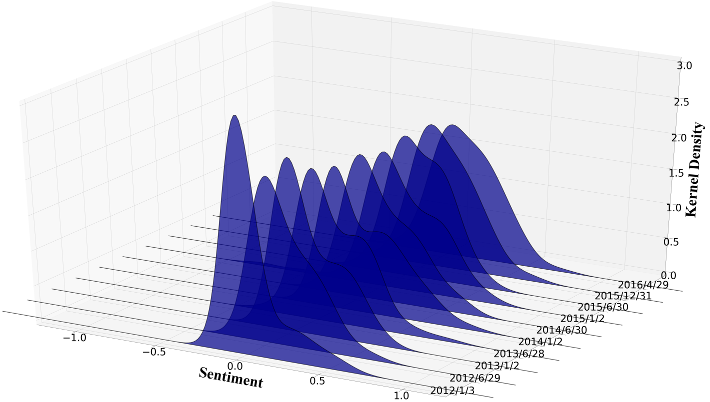
Source of the predictability ctd

- Sentiment-adjusted OCs remain significant
- Thus some information embedded in options markets data contains information other than sentiment
- Sentiment indices remain significant.
- Stock-specific bullishness not important.

Market consensus and stock returns

- data yield a cross section of firm-level sentiment measures
- observations are varying over time
- how does dispersion of sentiment affect stock returns?
  - low dispersion: cross-sectionally unequivocal sentiment
  - high dispersion: cross-sectionally differing sentiment
- implications unclear:
  - Miller (1977): dispersion could lead be negatively related to returns if pessimists stay out of the market due to short sale constraints
  - Varian (1985); Cujean and Hasler (2016): investors demand compensation, e.g. due to adverse selection.
- mesasure dispersion by cross-sectional standard deviation and include in predictive regressions

Cross-section of B_i



Stock return predictability: Option variables and sentiment

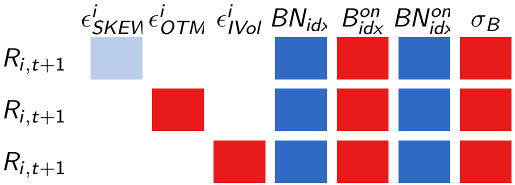


Table: Significance codes 0.01 0.05 0.1 0.01 0.05 0.1

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Blue (negative sign);Red (positive sign)

Market consensus and stock returns

- sentiment dispersion commands a high positive risk premium in the presence of market/ idiosyncratic volatility
- indeed sentiment dispersion and market volatility are only weakly correlated
- investors demand compensation for holding assets when sentiment is dispersed
- lends support to Varian (1985) / Cujean and Hasler (2016) among others

Trading

- Xing et al. (2010) show OC based trading strategies yield positive returns.
- Do OC stratgies after partialling out sentiment do better?
- Strategy:
  - Group data of 97 firms into deciles according to OC / OC residuals
  - create long-short portfolios on the extreme deciles.

Trading strategies						
	Skew residual			*Skew*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	14.42	14.74	14.77	14.18	14.61	14.58
P value	0.002	0.002	0.002	0.004	0.004	0.004
Ann. Return	0.43	0.45	0.45	0.43	0.44	0.44
Daily Vol. (in bp)	86.25			92.79		
Ann. Vol.	0.14			0.15		
Daily Sharpe Ratio	0.17			0.15		
Ann. Sharpe Ratio	3.18			2.91		
	IV residual			*IV*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	12.41	12.54	12.57	6.79	7.14	7.26
P value	0.009	0.010	0.010	0.181	0.121	0.141
Ann. Return	0.36	0.37	0.37	0.19	0.20	0.20
Daily Vol. (in bp)	88.67			99.28		
Ann. Vol.	0.14			0.16		
Daily Sharpe Ratio	0.14			0.07		
Ann. Sharpe Ratio	2.59			1.18		
	Put residual			*Put*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	7.43	7.86	7.70	6.52	6.92	6.87
P value	0.098	0.090	0.098	0.178	0.118	0.140
Ann. Return	0.20	0.22	0.21	0.18	0.19	0.19
Daily Vol. (in bp)	85.66			94.18		
Ann. Vol.	0.14			0.15		
Daily Sharpe Ratio	0.09			0.07		
Ann. Sharpe Ratio	1.51			1.19		

Summary

- We connect investor sentiment distilled from public news with equity and equity options markets
- Options markets react to firm-level sentiment and aggregate sentiment
- Relevance of inside information in option data after partialling out sentiment information from option data.
- Negative bullishness indices are important regressors in predictive regressions.
- Market consensus carries a positive risk premium.
- OC residual-based trading strategies slightly outperform pure OC based strategies.
- Results robust to lexicon projection techniques.

Textual Sentiment, Option Information and Stock Return Predictability

Cathy Yi-Hsuan Chen
Matthias R. Fengler
Wolfgang Karl Härdle
Yanchu Liu

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
University of St. Gallen, Switzerland
Lingnan College, Sun Yat-sen University, China

http://lvb.wiwi.hu-berlin.de
http://www.mathstat.unisg.ch/
http://www.lingnan.sysu.edu.cn

Bibliography

Antweiler, W. and Frank, M. Z.
Is All That Talk Just Noise?
J. Finance, 2004

Dennis, P. and S. Mayhew.
Risk-Neutral Skewness: Evidence from Stock Options
J. Financial Quant. Anal., 2002

Fama, E. and K. French.
A Five-Factor Asset Pricing Model
J. Financial Econom., 2015

Garman, M. and Klass, M.
On the Estimation of Security Price Volatilities from Historical Data
J. Bus., 1980

Han, B.
Investor Sentiment and Option Prices
Rev. Financial Stud., 2008

Härdle, W. K. and Lee, Y. J. and Schäfer D. and Yeh Y. R.
Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies
J. Forecast., 2009

Hu, M. and Liu, B.
Mining and Summarizing Customer Reviews
10th ACM SIGKDD, 2004

Loughran, T. and McDonald, B.
When is a liability not a liability?
J. Finance, 2011

Malo, Pekka and Sinha, Ankur and Korhonen, Pekka and Wallenius, Jyrki and Takala, Pyry
Good debt or bad debt
J. Assoc. Inf. Sci. Technol., 2014

Wilson, T. and Wiebe, J. and Hoffmann, P.
Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis
HLT-EMNLP, 2005

Xing, Y., X. Zhang and R. Zhao.
What does the Individual Option Volatility Smirk Tell Us about Future Equity Returns
J. Financial Quant. Anal., 2010

Zhang, J., Chen C. Y., Härdle, W. K. and Bommes, E.
Distillation of News into Analysis of Stock Reactions
J. Bus. Econom. Statist., 2016

Appendix

Correlation - Positive Sentiment

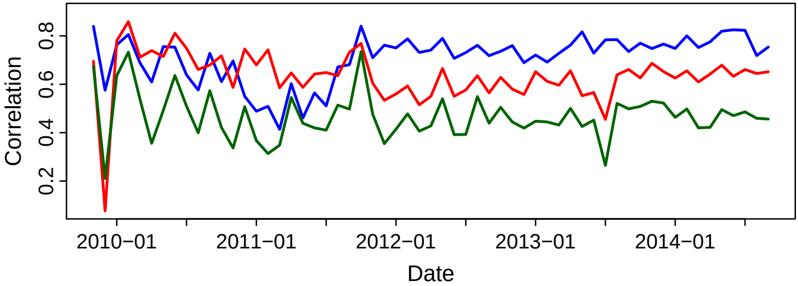


Figure: Monthly correlation between positive sentiment: BL and LM , BL and MPQA, LM and MPQA. Source: Zhang et al. (2016)

Back

Correlation - Negative Sentiment

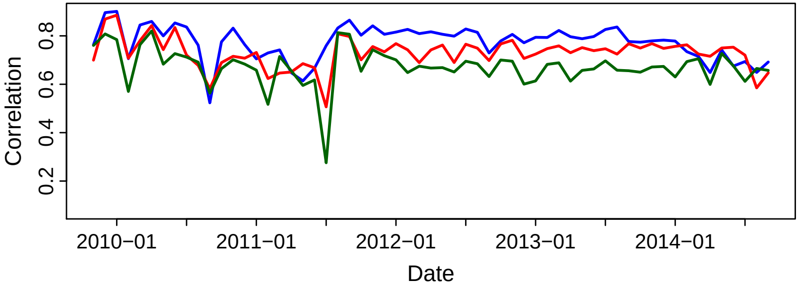


Figure: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2016) Back

Tagging Example - BL

... McDonald's has an obesity **problem** that continues to get **worse**. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem **like** a **good** thing, large menus result in **slower** service and more flare-ups between franchisees and the corporation.
Bloated menus raise inventory costs for smaller franchisees and **lead** to lower profit margins. The McDonald's corporate franchise fee is based upon sales instead of profits, making it a smaller **concern** for the company overall. ...

3 **positive words** and 5 **negative words**

TXTMcDbm
Article source

Tagging Example - LM

... McDonald's has an obesity **problem** that continues to get **worse**. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem like a **good** thing, large menus result in **slower** service and more flare-ups between franchisees and the corporation.
Bloated menus raise inventory costs for smaller franchisees and lead to lower profit margins. The McDonald's corporate franchise fee is based upon sales instead of profits, making it a smaller **concern** for the company overall. ...

1 **positive word** and 4 **negative words**

TXTMcDlm

Back

Web Scraping

- Databases to buy?
- Automatically extract information from web pages
- Transform unstructured data (HTML) to structured data
- Use HTML tree structure to parse web page
- Legal issues
  - Websites protected by copyright law
  - Prohibition of web scraping possible
  - Comply to Terms of Service (TOS)

Back

Natural Language Processing (NLP)

- Text is unstructured data with implicit structure
  - Text, sentences, words, characters
  - Nouns, verbs, adjectives, ..
  - Grammar
- Transform implicit text structure into explicit structure
- Reduce text variation for further analysis
- Python Natural Language Toolkit (NLTK)
- TXTnlp

Back

Tokenization

- String

  ```
  ''McDonald's has its work cut out for it.  Not only are
  sales falling in the U.S., but the company is now
  experiencing problems abroad.''
  ```

- Sentences

  ```
  ''McDonald's has its work cut out for it.'',
  ''Not only are sales falling in the U.S., but the company is
  now experiencing problems abroad.''
  ```

- Words

  ```
  ''McDonald'', '''s'', ''has'', ''its'', ''work'', ''cut'',
  ''out'' ...
  ```

Negation Handling

- "not good" $\neq$ "good"
- Reverse polarity of word if negation word is nearby
- Negation words
  `"n't", "not", "never", "no", "neither", "nor", "none"`

Part of Speech Tagging (POS)

- Grammatical tagging of words
  - `dogs - noun, plural (NNS)`
  - `saw - verb, past tense (VBD) or noun, singular (NN)`
- Penn Treebank POS tags
- Stochastic model or rule-based

Lemmatization

- Determine canonical form of word
  - `dogs - dog`
  - `saw (verb) - see and saw (noun) - saw`
- Reduces dimension of text
- Takes POS into account
  - Porter stemmer: `saw (verb and noun) - saw`

Back

Loss Functions for Classification

- Logistic: Logit

$$L\{y, s(X)\} = \log(2)^{-1} \log[1 + \exp\{-s(X)y\}] \qquad (4)$$

- Hinge: Support Vector Machines

$$L\{y, s(X)\} = \max\{0,\ 1 - s(X)y\} \qquad (5)$$

Back

Regularization Term

- L2 norm

$$R(\beta) = 2^{-1} \sum_{i=1}^{p} \beta_i^2 \tag{6}$$

- L1 norm

$$R(\beta) = \sum_{i=1}^{p} |\beta_i| \tag{7}$$

Back

RLM Example

Sentence 1: “The profit of Apple increased.”
Sentence 2: “The profit of the company decreased.”

$$y = (1, -1) \quad (8)$$

$$X = \begin{array}{c c} & \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{array}{c} \textit{the} \\ \textit{profit} \\ \textit{of} \\ \textit{Apple} \\ \textit{increased} \\ \textit{company} \\ \textit{decreased} \end{array} & \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \quad (9)$$

Back

k-fold Cross Validation (CV)

- Partition data into k complementary subsets
- No loss of information as in conventional validation
- Stratified CV: equally distributed response variable in each fold

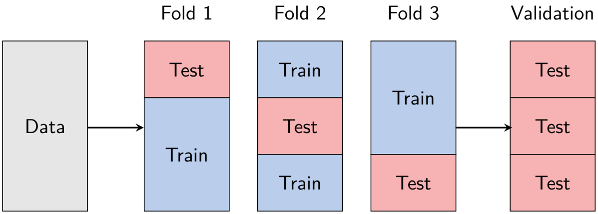


Figure: 3-fold Cross Validation

Back

Oversampling

- Härdle (2009) Trade-off between Type I and Type 2 error in classification Error types
- Balance size of neutral sentences and ones with polarity in sample
- Duplicate sentences within folds of stratified cross validation until the sample is balanced

Back

Classification Error Rates

- Type I error rate $= \text{FP}/(\text{FP} + \text{TP})$
- Type II error rate $= \text{FN}/(\text{FN} + \text{TP})$
- Total error rate $= (\text{FN} + \text{FP})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$

with TP as true positive, TN as true negative, FP as false positive and FN as false negative.

Back

Stochastic Gradient Descent (SGD)

- Approximately minimize loss function

$$L(\theta) = \sum_{i=1}^{n} L_i(\theta) \tag{10}$$

- Iteratively update

$$\theta_i = \theta_{i-1} - \eta \, \frac{\partial L_i(\theta)}{\partial \theta} \tag{11}$$

SGD Algorithm

1. Choose learning rate η
2. Shuffle data
3. For $i = 1, \ldots, n$, do:

$$\theta_i = \theta_{i-1} - \eta \, \frac{\partial L_i(\theta)}{\partial \theta}$$

Repeat 2 and 3 until approximate minimum obtained.

SGD Example

$X \sim \mathrm{N}(\mu, \sigma)$ and $x_1, ..., x_n$ as randomly drawn sample

$$\min_{\theta} \; n^{-1} \sum_{i=1}^{n} (\theta - x_i)^2$$

Update step

$$\theta_i = \theta_{i-1} - 2\eta(\theta_{i-1} - x_i)$$

Optimal gain

Set $2\eta = 1/i$ and obtain $\theta_n = \bar{x}$ with $\bar{x}$ as sample mean.

SGD Example ctd

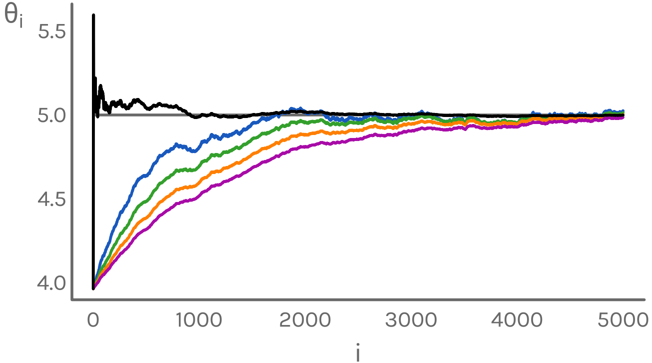


Figure: Estimate Mean via SGD, $x_t \sim N(5, 1)$

$\eta \in \{1/t,\ 1/1000,\ 1/1500,\ 1/2000,\ 1/2500\}$ TXTSGD

Back

Evaluation Supervised Learning

True \ Pred	-1	0	1	Total
-1	**1,992**	289	254	2,535
0	96	**2,134**	305	2,535
1	105	469	**1,961**	2,535
Total	2,193	2,892	2,520	7,605
Precision	0.91	0.74	0.78	
Recall	0.78	0.84	0.77	

Table: Confusion Matrix - Supervised Learning with Oversampling

Evaluation Unsupervised Learning

True \ Pred	-1	0	1	Total
-1	**213**	289	12	514
0	200	**2,187**	148	2,535
1	111	772	**285**	1,168
Total	524	3,248	445	4,217
Precision	0.41	0.67	0.64	
Recall	0.41	0.86	0.24	

Table: Confusion Matrix - Lexicon Projection

LDA – details

Assumed process of generating a document:

1. Choose number of words N (randomly, deterministically).
2. Draw a distribution over K topics:

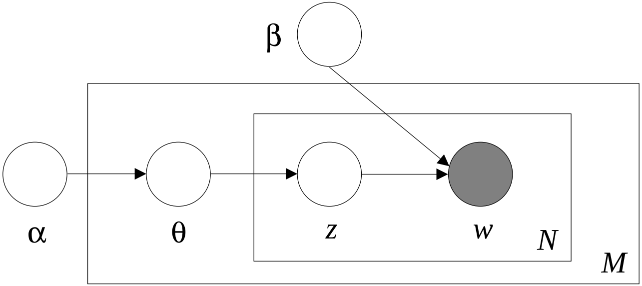
$$\theta \sim \text{Dir}(\alpha)$$

3. For each of the N words w_n:

   3.1 Choose a topic from $z_n \sim \text{M}(\theta)$

   3.2 Choose a word from $p(w_n|z_n, \beta)$, a multinomial probability conditional on topic z_n parametrized by $\beta = [\beta_{ij}] = p(w^j = 1|z^i = 1)$

Graphical representation of the LDA



Source: Blei et al. (2003)

Inference

- The estimation problem is to find the hidden topic structure over the set of documents given observed words.
- Need to approximate the posterior distribution, i.e., the conditional distribution of topics, topic proportions, and topic assignments given observed words.
- Posterior computation is achieved by Gibbs sampling, see Blei et al. (2012) for details.

Back

A plot of Skew

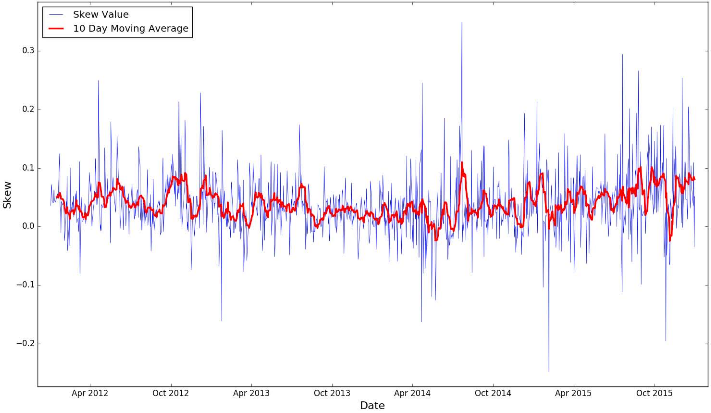


Figure: Skew of Apple Inc. in the sample period

Control Variables

Ret_{it} - Stock i's contemporous return

$Volu_{it}$ - Stock i's trading volume

OC_{it} - option characteristics of stock i

VIX_t - CBOE VIX More Information

and Fama-French 5 factors (Fama and French (JFE, 2015))

More Information

Back

Fama-French 5 factors

$FF1$ - the Mkt factor: excess return on the market index

$FF2$ - the SMB factor: (Small Minus Big) the average return on the nine small-stock portfolios minus that on the nine big-stock portfolios.

$FF3$ - the HML factor: (High Minus Low) the average return on the two value-stock portfolios minus that on the two growth-stock portfolios

Fama-French 5 factors ctd

$FF4$ - the RMW factor: (Robust Minus Weak) the average return on the two robust operating profitability portfolios minus that on the two weak operating profitability portfolios

$FF5$ - the CMA factor: (Conservative Minus Aggressive) the average return on the two conservative investment portfolios minus that on the two aggressive investment portfolios

VIX

- Implied volatility
- Measures market expectation of S&P 500
- Calculated by Chicago Board Options Exchange (CBOE)
- Measures 30-day expected volatility
- Calculated with put and call options with more than 23 days and less than 37 days to expiration

Back

Variables Definitions

- *Skew*: difference between volume-weighted average of implied volatilities (IVs) of OTMP and ATMC:

$$SKEW_{it} = IV_{it}^{OTMP} - IV_{it}^{ATMC}$$

Example

- *OTMP*: a put with moneyness between 0.8 and 0.95
- *ATMC*: a call with moneyness between 0.95 and 1.05
- Moneyness: ratio of the strike price to the stock price
- Use delta as moneyness

Variables Definitions ctd

- *IVol*: volume-weighted average of IVs of all the ATM options
- *OTM*: volume-weighted average of prices of OTM put options (moneyness between 0.8 and 0.95) *relative* to stock price
- *B*: degree of bullishness defined in (4), positive (negative) value implies positive (negative) sentiment
- $BN = -\mathbf{I}\left(B < 0\right)B$, indicating negative sentiment

Back