GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Prof. Dr. Alexander Hillert

Textual Analysis

Chapter 2: Literature Review and Methodology

July 18, 20, and 22, 2022

WHU – Otto Beisheim School of Management

Agenda

- Main contributions on textual analysis in finance
  - Naïve Bayes approach: Antweiler and Frank (2004)
  - Introduction to machine learning
  - Dictionary approach: Tetlock (2007) and Loughran and McDonald (2011)
- Recommendations for your textual analysis
- Selected topics and papers
  - Readability: Loughran and McDonald (2014, 2020)
  - Textual similarity: Tetlock (2011) and Cohen et al. (2020)

1. Antweiler and Frank (2004)

Antweiler and Frank (2004) – Motivation (1)

Motivation for the paper

- During the late 1990s and early 2000s stock message boards were very popular.
- The financial press claimed that stock message boards can influence prices.
- There is anecdotal evidence for that:
  - Main characters: Arash Aziz-Golshani (23), Hootan Melamed (23), Allen Derzakharian (26).
  - Date: October and November 1999.
  - What happened on the stock market?
    - Friday November 12, NEI Webworld Inc.’s stock closed at 13 cents.
    - Monday November 15, it opened at 8 dollar.
    - After half an hours of trading the NEI’s price was 15.50 dollar.
    - NEI shares closed at 75 cents.
    - December 15: stock closed 18.75 cents.
  - What caused the large stock price changes?

Antweiler and Frank (2004) – Motivation (2)

Introductory example – continued

- The three people
  - Bought 97% of shares of NEI Webworld Inc. at prices between 5 cents and 17 cents per share.
  - Then used public computers in the biomedical library at the UCLA to promote NEI Webworld Inc.
    - from 50 accounts they posted more than 500 messages on three websites.
    - Example: "Buying NEIP early would entitle you to a share of LGC Wireless when it goes public next week. Look for a massive move to $5-$10 as wireless stocks are very hot."
  - Sold their shares into the buying frenzy at prices ranging from 25 cents to $15.18.
  - Payoffs:
    - Arash Aziz-Golshani: $152,742
    - Hootan Melamed and Allen Derzakharian (joint brokerage account): $211,250.

Antweiler and Frank (2004) – Motivation (3)

Definition of market manipulation:

- “**Manipulation is intentional conduct designed to deceive investors by controlling or artificially affecting the market for a security**. Manipulation can involve a number of techniques to affect the supply of, or demand for, a stock. They include: **spreading false or misleading information about a company**; improperly limiting the number of publicly-available shares; or rigging quotes, prices or trades to create a false or deceptive picture of the demand for a security. Those who engage in manipulation are subject to various civil and criminal sanctions.”
  Source: https://www.sec.gov/fast-answers/answerstmanipulhtm.html, posted March 28, 2008.

Antweiler and Frank (2004) – Motivation (4)

Introductory example – continued

- Legal consequences:
  - Arash Aziz-Golshani: sentenced to 15 months incarceration on January 22, 2001.
  - Hootan Melamed: sentenced to 10 months incarceration on January 12, 2001.
  - Allen Derzakharian: civil fraud but not charged in the criminal case.
  - In addition: pay restitution

- Aziz-Golshani and Melamed engaged in security fraud earlier!

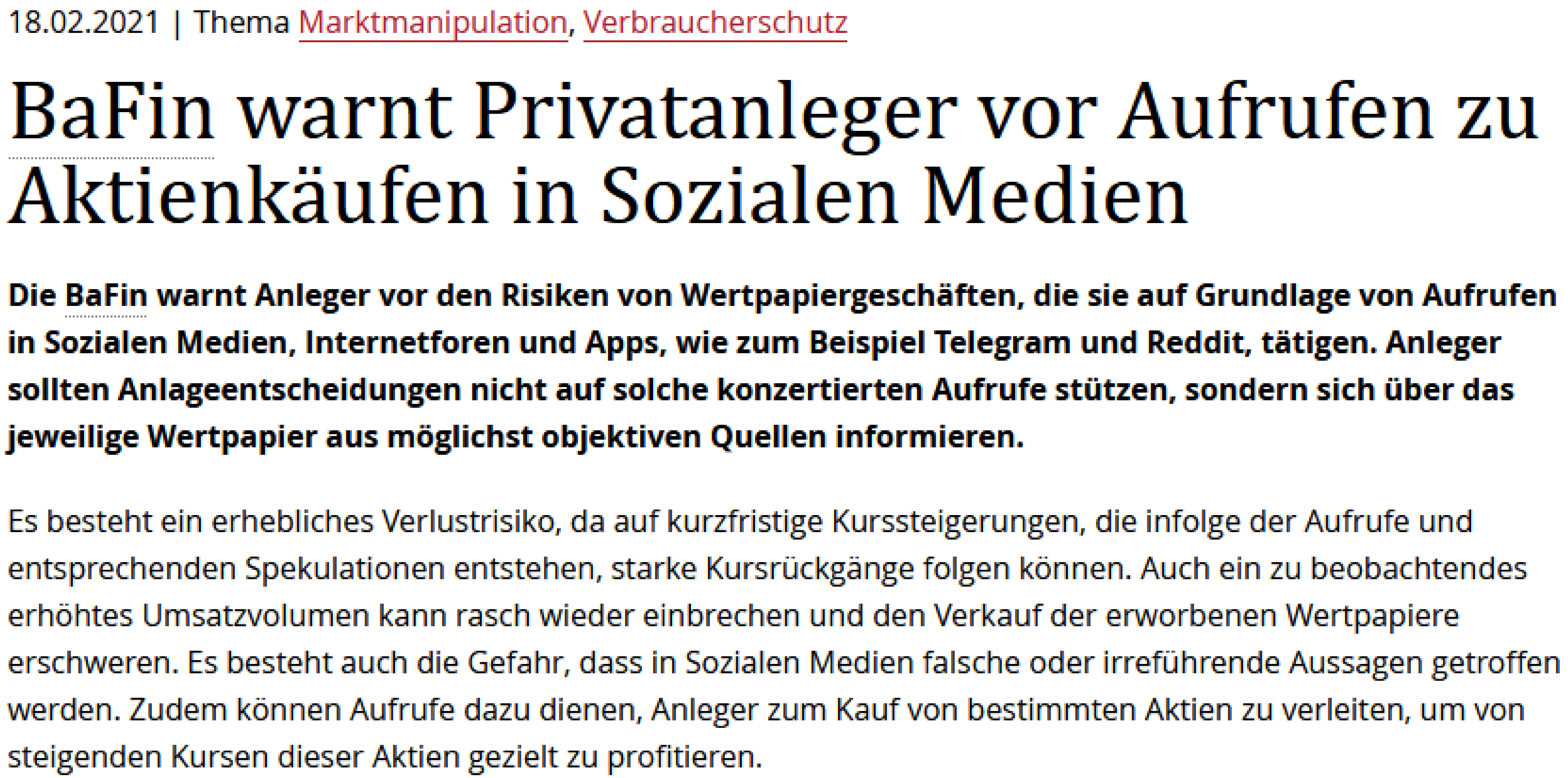
Antweiler and Frank (2004) – Motivation (5)

Warning by the SEC regarding stock message boards

“Online bulletin boards—whether newsgroups, blogs, or web-based bulletin boards—have become an increasingly popular forum for investors to share information. Bulletin boards typically feature "threads" made up of numerous messages on various investment opportunities.
While some messages may be true, many turn out to be bogus—or even scams. Fraudsters often pump up a company or pretend to reveal "inside" information about upcoming announcements, new products, or lucrative contracts.
Also, you never know for certain who you’re dealing with—or whether they’re credible—because many bulletin boards allow users to hide their identity behind multiple aliases. People claiming to be unbiased observers who’ve carefully researched the company may actually be company insiders, large shareholders, or paid promoters. A single person can easily create the illusion of widespread interest in a small, thinly traded stock by posting a series of messages under various aliases.”
Source: https://www.sec.gov/fast-answers/answersmsgbdhtm.html, posted July 5, 2000

Antweiler and Frank (2004) – Motivation (6)

Recent example of the effects of discussion websites: GameStop



18.02.2021 | Thema Marktmanipulation, Verbraucherschutz

BaFin warnt Privatanleger vor Aufrufen zu Aktienkäufen in Sozialen Medien

Die BaFin warnt Anleger vor den Risiken von Wertpapiergeschäften, die sie auf Grundlage von Aufrufen in Sozialen Medien, Internetforen und Apps, wie zum Beispiel Telegram und Reddit, tätigen. Anleger sollten Anlageentscheidungen nicht auf solche konzertierten Aufrufe stützen, sondern sich über das jeweilige Wertpapier aus möglichst objektiven Quellen informieren.

Es besteht ein erhebliches Verlustrisiko, da auf kurzfristige Kurssteigerungen, die infolge der Aufrufe und entsprechenden Spekulationen entstehen, starke Kursrückgänge folgen können. Auch ein zu beobachtendes erhöhtes Umsatzvolumen kann rasch wieder einbrechen und den Verkauf der erworbenen Wertpapiere erschweren. Es besteht auch die Gefahr, dass in Sozialen Medien falsche oder irreführende Aussagen getroffen werden. Zudem können Aufrufe dazu dienen, Anleger zum Kauf von bestimmten Aktien zu verleiten, um von steigenden Kursen dieser Aktien gezielt zu profitieren.

We will talk more about Reddit and GameStop later (Slides 28 to 34).

Source: https://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Verbrauchermitteilung/weitere/2021/meldung_210218_Warnung_vor_Aufrufen_zu_Aktienkaeufen.html, posted February 18, 2021

Antweiler and Frank (2004) – Research question and variables of interest

Antweiler and Frank (2004)

- Research question: do stock message boards contain information?

- Variables of Interest
  - Antweiler and Frank (AF) analyze
    - Stock prices
    - Trading volume (different order sizes)
    - Volatility
  - and measure
    - # messages posted
    - Bullishness (tone)
    - Agreement

Antweiler and Frank (2004) - Sample

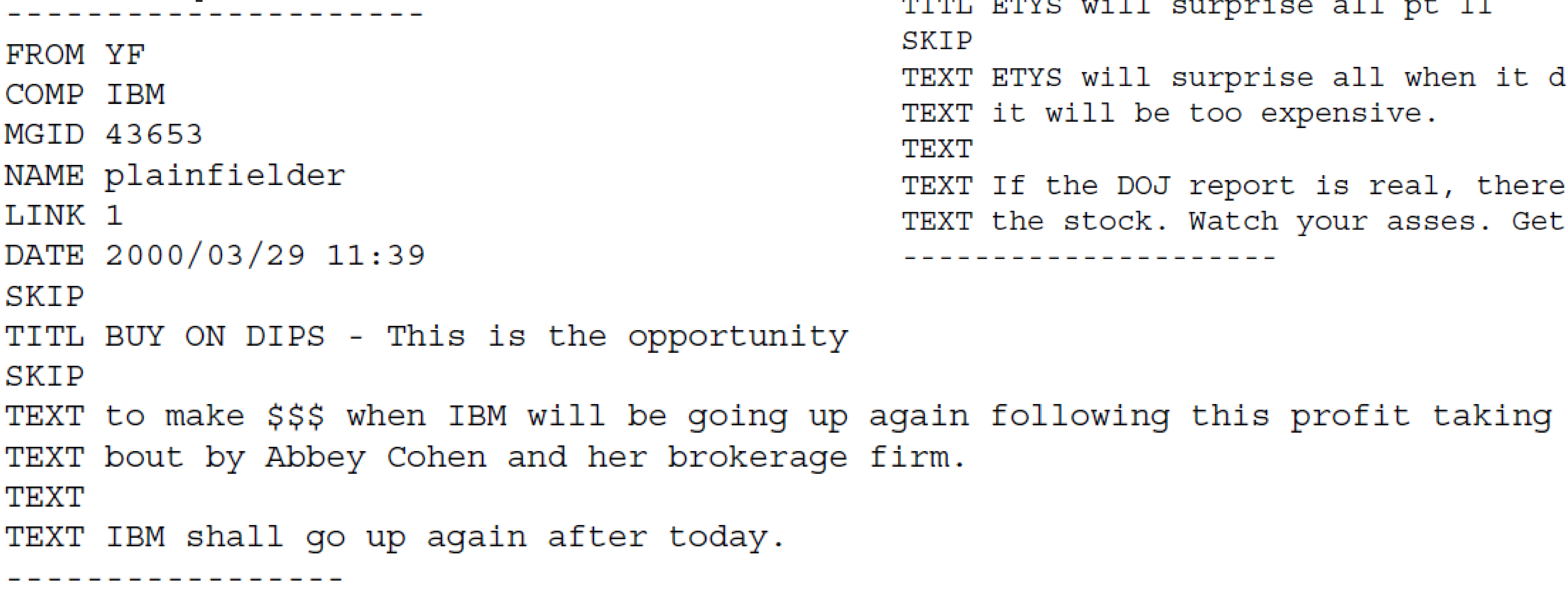
Data source

- Yahoo! Finance and Raging Bull
- “messages were downloaded using specialized software written by the authors” (AF, p.1262)
- Sample
  - ○ 30 stocks from the Dow Jones Industrial Average (DIA)
  - ○ 15 stocks from the Dow Jones Internet Commerce Index (XLK)
  - ○ January 1, 2000 to December 31, 2000.
  - ○ 1,560,621 messages.
    However, final sample much smaller (~331,000).

Antweiler and Frank (2004) – Examples

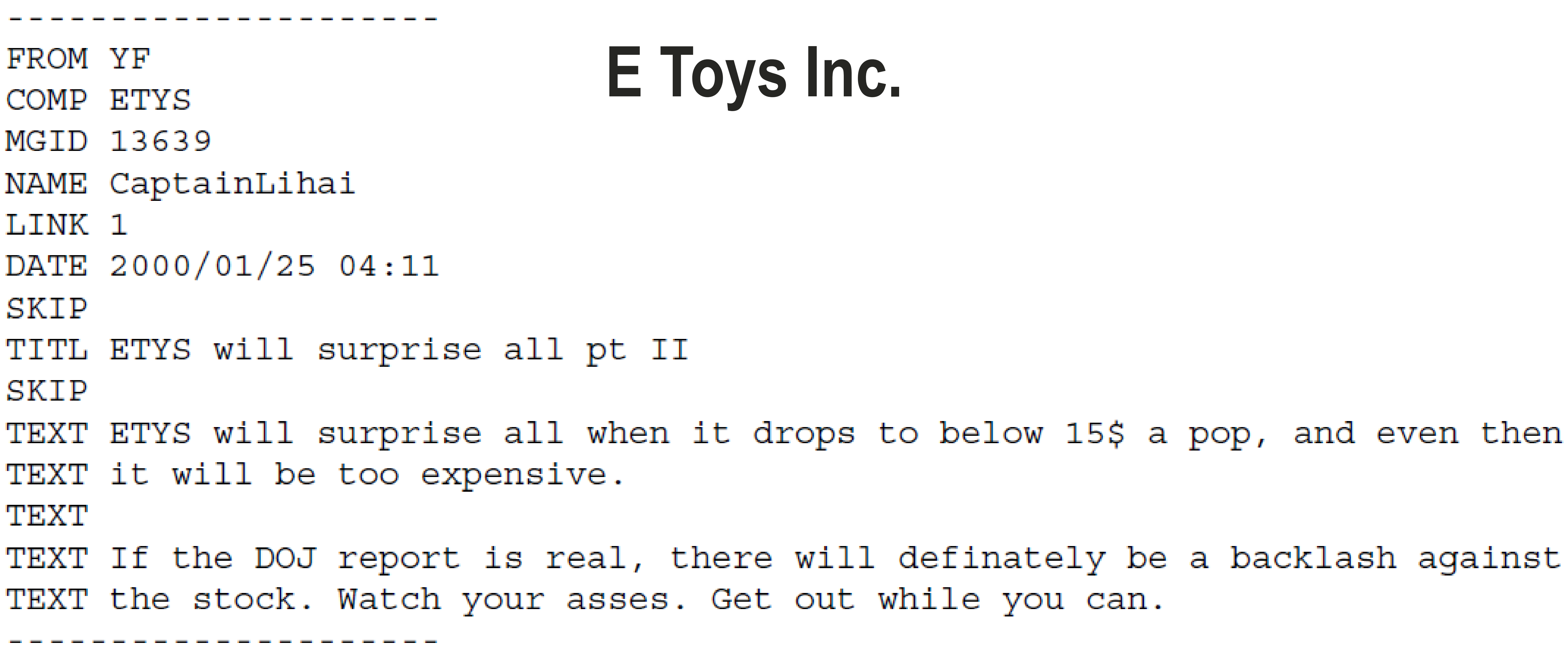
Examples

IBM Corp.



```
---------------------
FROM YF
COMP IBM
MGID 43653
NAME plainfielder
LINK 1
DATE 2000/03/29 11:39
SKIP
TITL BUY ON DIPS - This is the opportunity
SKIP
TEXT to make $$$ when IBM will be going up again following this profit taking
TEXT bout by Abbey Cohen and her brokerage firm.
TEXT
TEXT IBM shall go up again after today.
-----------------
```

E Toys Inc.



```
---------------------
FROM YF
COMP ETYS
MGID 13639
NAME CaptainLihai
LINK 1
DATE 2000/01/25 04:11
SKIP
TITL ETYS will surprise all pt II
SKIP
TEXT ETYS will surprise all when it drops to below 15$ a pop, and even then
TEXT it will be too expensive.
TEXT
TEXT If the DOJ report is real, there will definately be a backlash against
TEXT the stock. Watch your asses. Get out while you can.
---------------------
```

AF have information on company, unique message ID, author of the message, date (minute), and text.

Antweiler and Frank (2004) – Tone measurement (1)

How to measure tone?

AF measure the tone by Naïve Bayes

- Why ‘naïve’?
  → ignores the grammatical structure of texts.

Procedure

- Randomly choose a set of training documents. AF use 1,000 messages.
- Researchers/people classify all messages. AF use three categories (positive, neutral, negative).
- Feed textual analysis software with the training data.
  → AF use the Rainbow package developed by McCallum (1996).
  → We will apply Naïve Bayes using Python’s NLTK package.
- Software will classify the remaining sample based on the training data.
  → How does this work?

Antweiler and Frank (2004) – Tone measurement (2)

Variables:

- T : type (positive, neutral, negative)
- $\tilde{T}$: anti-type, (not positive, not neutral, not negative)
- W_i: i^{th} word in the text
- $P(T|W_0)=P(T)$ unconditional probability

Bayes' theorem $P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$

$$P(T|W_i) = \frac{\overbrace{P(T|W_{i-1}) \cdot P(W_i|T)}^{\text{How likely is it to observe word } W_i \text{ and the text being of type T?}}}{\underbrace{P(T|W_{i-1}) \cdot P(W_i|T) + \left(1 - P(T|W_{i-1})\right) \cdot P(W_i|\tilde{T})}_{\text{How likely is it to observe word } W_i \text{ in any type of text?}}}$$

- Intuitively, we start with the unconditional probability for pos/neu/neg articles and update our best estimate for the document being a certain type with each word in the document.

Antweiler and Frank (2004) – Tone measurement (3)

Example:

- From the training dataset: unconditional probability of a negative article is 25%.
- The first word of the message is ‘loss’.
- The word ‘loss’ occurs in 70% of negative articles but only in 20% of neutral and positive articles.

You get this information from the training data set, i.e. from the manual classification of articles.

Unconditional Probability P(T|W(0)) =	0.25
First word of article W(1) is "Loss"	
P("Loss"|Neg) =	0.7
P("Loss"|Not Neg) =	0.2
P(T|W(1)) = P(Neg|"Loss") =	0.25*0.7 / (0.25*0.7 + (1-0.25)*0.2)
=> P(Neg|W(1)) =	0.5385

$$P(T|W_i) = \frac{P(T|W_{i-1}) \cdot P(W_i|T)}{P(T|W_{i-1}) \cdot P(W_i|T) + \left(1 - P(T|W_{i-1})\right) \cdot P(W_i|\tilde{T})}$$

Antweiler and Frank (2004) – Tone measurement (4)

Example - continued

- P(Neg|W(1)) = 53.85%
- The second word of the message is ‘benefit’.
- The word ‘benefit’ occurs in only 10% of negative articles but in 75% of neutral and positive articles.

Second word of article W(2) is "benefit"	
P("benefit"|Neg) =	0.1
P("benefit"|Not Neg) =	0.75
P(T|W(2)) = P(Neg|"benefit") =	0.5385*0.1 / (0.5385*0.1 + (1-0.5385)*0.75)
=> P(Neg|W(2)) =	0.1346

$$P(T|W_i) == \frac{P(T|W_{i-1}) \cdot P(W_i|T)}{P(T|W_{i-1}) \cdot P(W_i|T) + \left(1 - P(T|W_{i-1})\right) \cdot P(W_i|\tilde{T})}$$

Antweiler and Frank (2004) – Tone measurement (5)

Example - continued

- P(Neg|W(2)) = 13.46%
- The third word of the message is ‘sales’.
- The word ‘sales’ occurs in 80% of negative, positive, and neutral articles.

Third word of article W(3) is "sales"	
P("sales"|Neg) =	0.8
P("sales"|Not Neg) =	0.8
P(T|W(3)) = P(Neg|“sales“) =	0.1346*0.8 / (0.1346*0.8 + (1-0.1346)*0.8)
=> P(Neg|W(3)) =	0.1346

$$P(T|W_i) == \frac{P(T|W_{i-1}) \cdot P(W_i|T)}{P(T|W_{i-1}) \cdot P(W_i|T) + \left(1 - P(T|W_{i-1})\right) \cdot P(W_i|\tilde{T})}$$

Calculation of tone in practice:

$$P(T|W_N) = P(T)\exp\left[\sum_{i=1}^{N} \ln\left(\frac{P(W_i|T)}{P(W_i|\tilde{T})}\right)\right]$$

- N is the number of words.
- If word i is more likely to occur in type T than in the anti-type then $P(W_i|T) > P(W_i|\tilde{T})$ → ln() > 0 → increase in the probability of the message being type T.

Antweiler and Frank (2004) – Tone measurement (7)

Performance of classification algorithm:

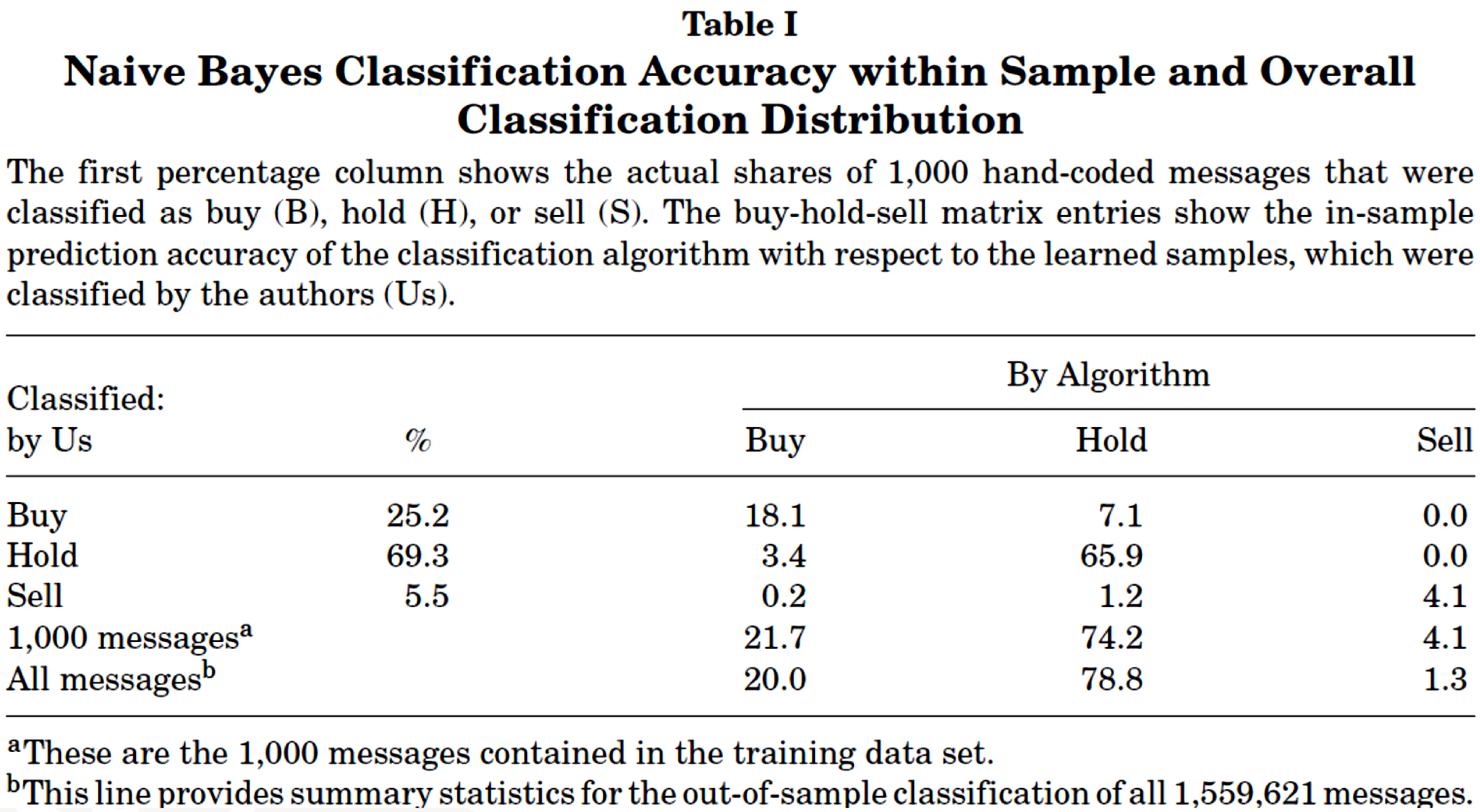


Table I

Naive Bayes Classification Accuracy within Sample and Overall Classification Distribution

The first percentage column shows the actual shares of 1,000 hand-coded messages that were classified as buy (B), hold (H), or sell (S). The buy-hold-sell matrix entries show the in-sample prediction accuracy of the classification algorithm with respect to the learned samples, which were classified by the authors (Us).

Classified:		By Algorithm		
by Us	%	Buy	Hold	Sell
Buy	25.2	18.1	7.1	0.0
Hold	69.3	3.4	65.9	0.0
Sell	5.5	0.2	1.2	4.1
1,000 messages[a]		21.7	74.2	4.1
All messages[b]		20.0	78.8	1.3

[a] These are the 1,000 messages contained in the training data set.

[b] This line provides summary statistics for the out-of-sample classification of all 1,559,621 messages.

Comments:

- The numbers are frequencies relative to the training/testing sample. Example: percentage of correctly identified buys = 18.1 / 25.2 = 71.83%
- There is no information on out-of-sample accuracy. → A second set of manually classified articles for validating the out-of-sample classification would be helpful.
- As training set is randomly drawn from overall population, we see that about 20% of buys (5.2/25.2) and about 76% of sells (4.2/5.5) are likely to be misclassified.

Antweiler and Frank (2004) – Tone measurement (8)

Aggregation of tone:

How to aggregate message tone into a single bullishness indicator?

- AF define:
  - ○ $M_t = M_t^{Buy} + M_t^{Sell}$
  - ○ $R_t = M_t^{Buy} / M_t^{Sell}$
- First measure:

$$B_t \equiv \frac{M_t^{BUY} - M_t^{SELL}}{M_t^{BUY} + M_t^{SELL}} = \frac{R_t - 1}{R_t + 1}$$

  - ○ Independent of the number of messages
  - ○ Bound between -1 and 1
- Second measure:

$$B_t^* \equiv \ln\left[\frac{1 + M_t^{BUY}}{1 + M_t^{SELL}}\right] = \ln\left[\frac{2 + M_t(1 + B_t)}{2 + M_t(1 - B_t)}\right] \approx B_t \ln(1 + M_t)$$

Accounts for the number of messages

Antweiler and Frank (2004) – Tone measurement (9)

Aggregation of tone - continued

- AF define:
  - ○ $M_t = M_t^{Buy} + M_t^{Sell}$
  - ○ $R_t = M_t^{Buy} / M_t^{Sell}$
- Third measure:

$$B_t^{**} \equiv M_t^{BUY} - M_t^{SELL} = M_t \left[\frac{R_t - 1}{R_t + 1}\right] = M_t B_t$$

→ Increases linearly in the number of messages.

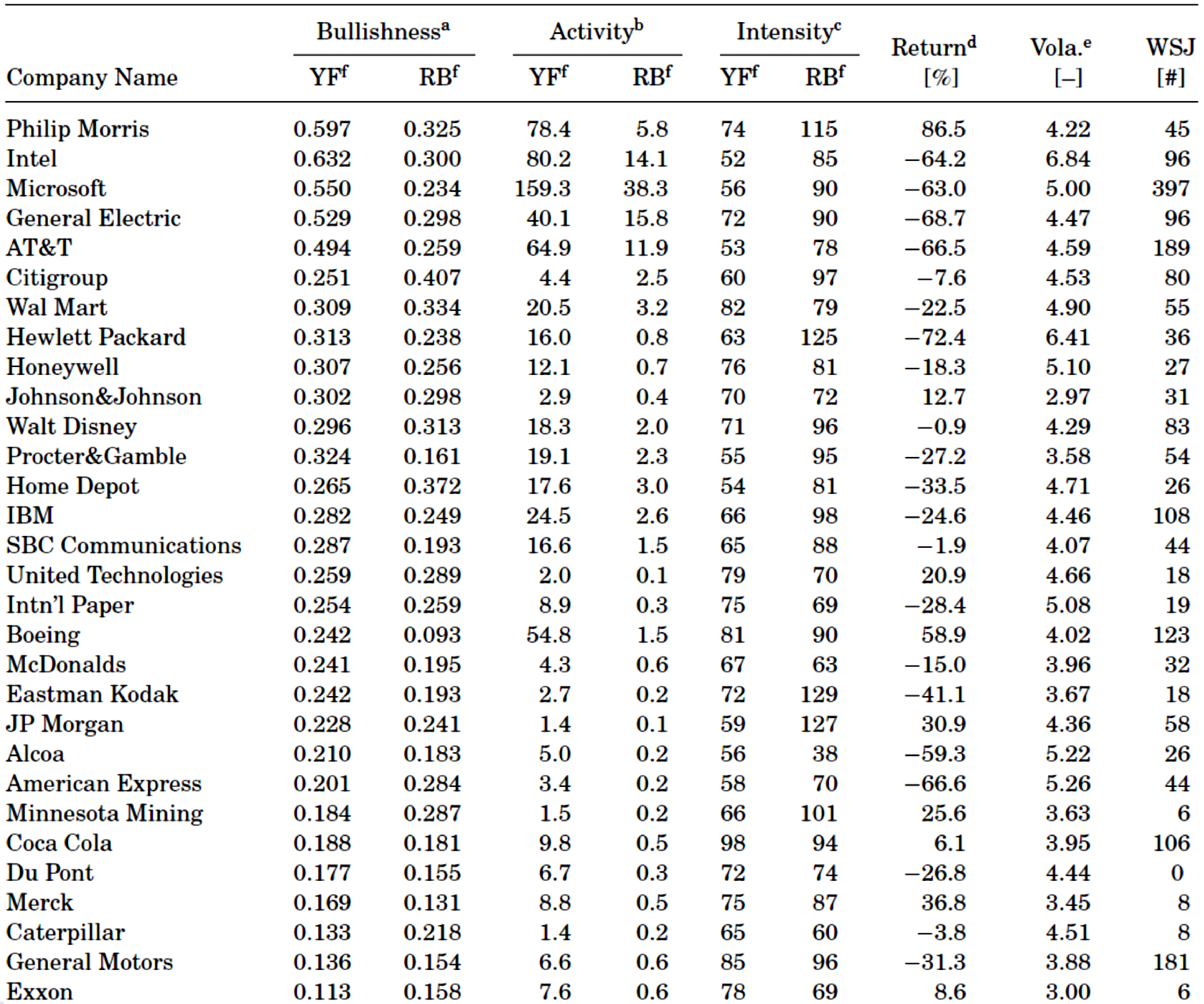
- AF use the second measure B*.
  "The measure B* appears to outperform both alternatives and so we use it in all reported tables.“ (p. 1267)
- AF do not include neutral messages.
  Footnote 11: ‘In none of our three measures do we use the number of “hold” messages. This group contains both “noise” as well as neutral (hold) opinions. The amount of “noise” dominates.’
  Note that this excludes 78.8% of all messages! → final sample only 331,000 observations.
- AF assume that B = 0 for periods with zero messages.

Antweiler and Frank (2004) – Descriptives (1)

Summary statistics of the Dow Jones 30 companies – AF, Table 2

- Table is sorted by average bullishness.
- Internet-/telecommunication-related companies
  - receive more attention ("Activity" = messages in 1,000s).
  - are discussed more favorably despite poor stock returns.

- Companies covered in stock message boards are different from those covered by the Wall Street Journal (last column).

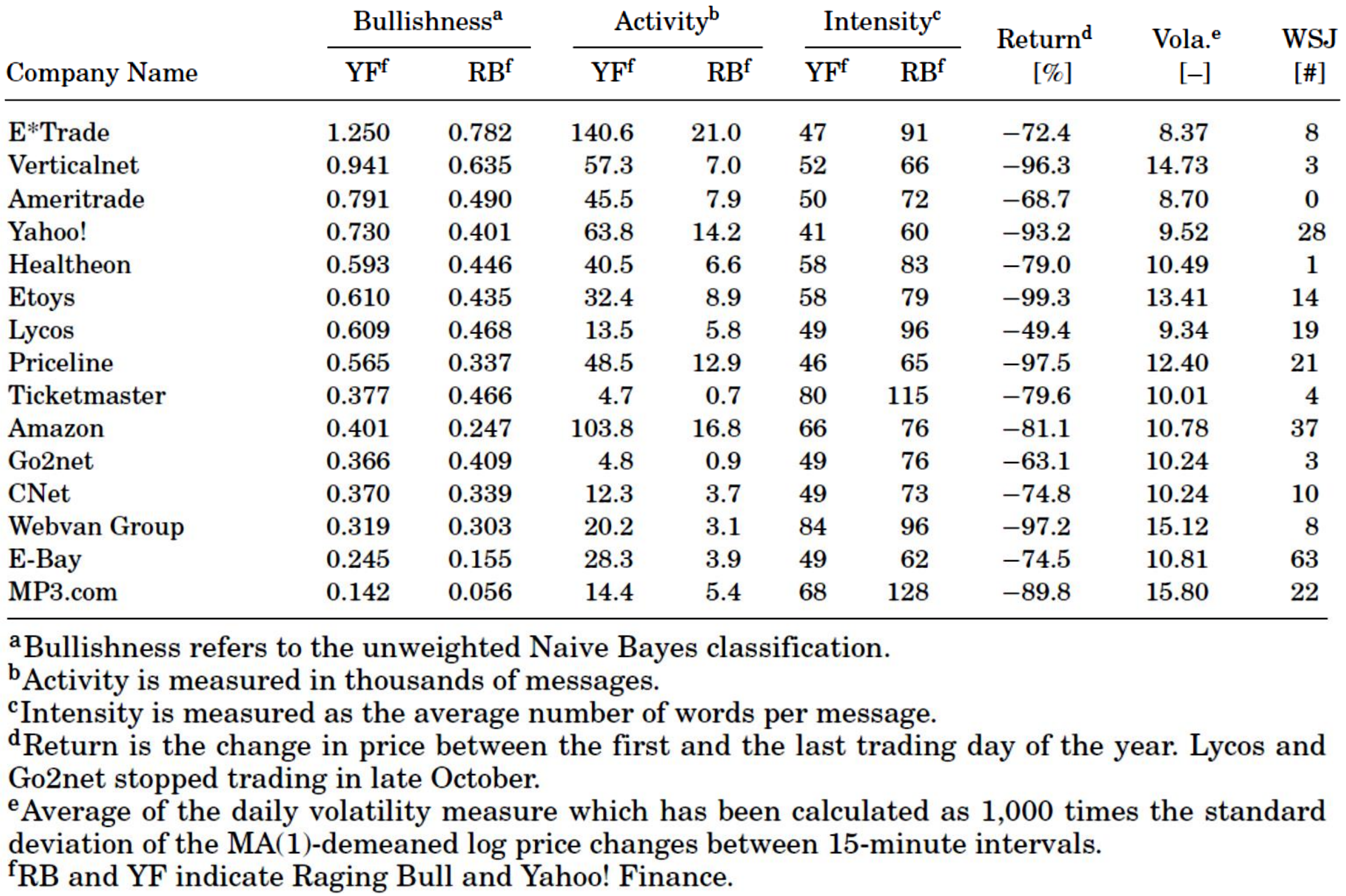


	Bullishness[a]		Activity[b]		Intensity[c]		Return[d]	Vola.[e]	WSJ
Company Name	YF[f]	RB[f]	YF[f]	RB[f]	YF[f]	RB[f]	[%]	[–]	[#]
Philip Morris	0.597	0.325	78.4	5.8	74	115	86.5	4.22	45
Intel	0.632	0.300	80.2	14.1	52	85	−64.2	6.84	96
Microsoft	0.550	0.234	159.3	38.3	56	90	−63.0	5.00	397
General Electric	0.529	0.298	40.1	15.8	72	90	−68.7	4.47	96
AT&T	0.494	0.259	64.9	11.9	53	78	−66.5	4.59	189
Citigroup	0.251	0.407	4.4	2.5	60	97	−7.6	4.53	80
Wal Mart	0.309	0.334	20.5	3.2	82	79	−22.5	4.90	55
Hewlett Packard	0.313	0.238	16.0	0.8	63	125	−72.4	6.41	36
Honeywell	0.307	0.256	12.1	0.7	76	81	−18.3	5.10	27
Johnson&Johnson	0.302	0.298	2.9	0.4	70	72	12.7	2.97	31
Walt Disney	0.296	0.313	18.3	2.0	71	96	−0.9	4.29	83
Procter&Gamble	0.324	0.161	19.1	2.3	55	95	−27.2	3.58	54
Home Depot	0.265	0.372	17.6	3.0	54	81	−33.5	4.71	26
IBM	0.282	0.249	24.5	2.6	66	98	−24.6	4.46	108
SBC Communications	0.287	0.193	16.6	1.5	65	88	−1.9	4.07	44
United Technologies	0.259	0.289	2.0	0.1	79	70	20.9	4.66	18
Intn'l Paper	0.254	0.259	8.9	0.3	75	69	−28.4	5.08	19
Boeing	0.242	0.093	54.8	1.5	81	90	58.9	4.02	123
McDonalds	0.241	0.195	4.3	0.6	67	63	−15.0	3.96	32
Eastman Kodak	0.242	0.193	2.7	0.2	72	129	−41.1	3.67	18
JP Morgan	0.228	0.241	1.4	0.1	59	127	30.9	4.36	58
Alcoa	0.210	0.183	5.0	0.2	56	38	−59.3	5.22	26
American Express	0.201	0.284	3.4	0.2	58	70	−66.6	5.26	44
Minnesota Mining	0.184	0.287	1.5	0.2	66	101	25.6	3.63	6
Coca Cola	0.188	0.181	9.8	0.5	98	94	6.1	3.95	106
Du Pont	0.177	0.155	6.7	0.3	72	74	−26.8	4.44	0
Merck	0.169	0.131	8.8	0.5	75	87	36.8	3.45	8
Caterpillar	0.133	0.218	1.4	0.2	65	60	−3.8	4.51	8
General Motors	0.136	0.154	6.6	0.6	85	96	−31.3	3.88	181
Exxon	0.113	0.158	7.6	0.6	78	69	8.6	3.00	6

Antweiler and Frank (2004) – Descriptives (2)

Summary statistics of the 15 telecommunication index companies – AF, Table 2

- Numbers confirm previous insights.
- Telecommunication companies
  - receive much more attention despite being smaller.
  - are discussed more favorably despite poor stock returns.



	Bullishness[a]		Activity[b]		Intensity[c]		Return[d]	Vola.[e]	WSJ
Company Name	YF[f]	RB[f]	YF[f]	RB[f]	YF[f]	RB[f]	[%]	[–]	[#]
E*Trade	1.250	0.782	140.6	21.0	47	91	−72.4	8.37	8
Verticalnet	0.941	0.635	57.3	7.0	52	66	−96.3	14.73	3
Ameritrade	0.791	0.490	45.5	7.9	50	72	−68.7	8.70	0
Yahoo!	0.730	0.401	63.8	14.2	41	60	−93.2	9.52	28
Healtheon	0.593	0.446	40.5	6.6	58	83	−79.0	10.49	1
Etoys	0.610	0.435	32.4	8.9	58	79	−99.3	13.41	14
Lycos	0.609	0.468	13.5	5.8	49	96	−49.4	9.34	19
Priceline	0.565	0.337	48.5	12.9	46	65	−97.5	12.40	21
Ticketmaster	0.377	0.466	4.7	0.7	80	115	−79.6	10.01	4
Amazon	0.401	0.247	103.8	16.8	66	76	−81.1	10.78	37
Go2net	0.366	0.409	4.8	0.9	49	76	−63.1	10.24	3
CNet	0.370	0.339	12.3	3.7	49	73	−74.8	10.24	10
Webvan Group	0.319	0.303	20.2	3.1	84	96	−97.2	15.12	8
E-Bay	0.245	0.155	28.3	3.9	49	62	−74.5	10.81	63
MP3.com	0.142	0.056	14.4	5.4	68	128	−89.8	15.80	22

[a] Bullishness refers to the unweighted Naive Bayes classification.

[b] Activity is measured in thousands of messages.

[c] Intensity is measured as the average number of words per message.

[d] Return is the change in price between the first and the last trading day of the year. Lycos and Go2net stopped trading in late October.

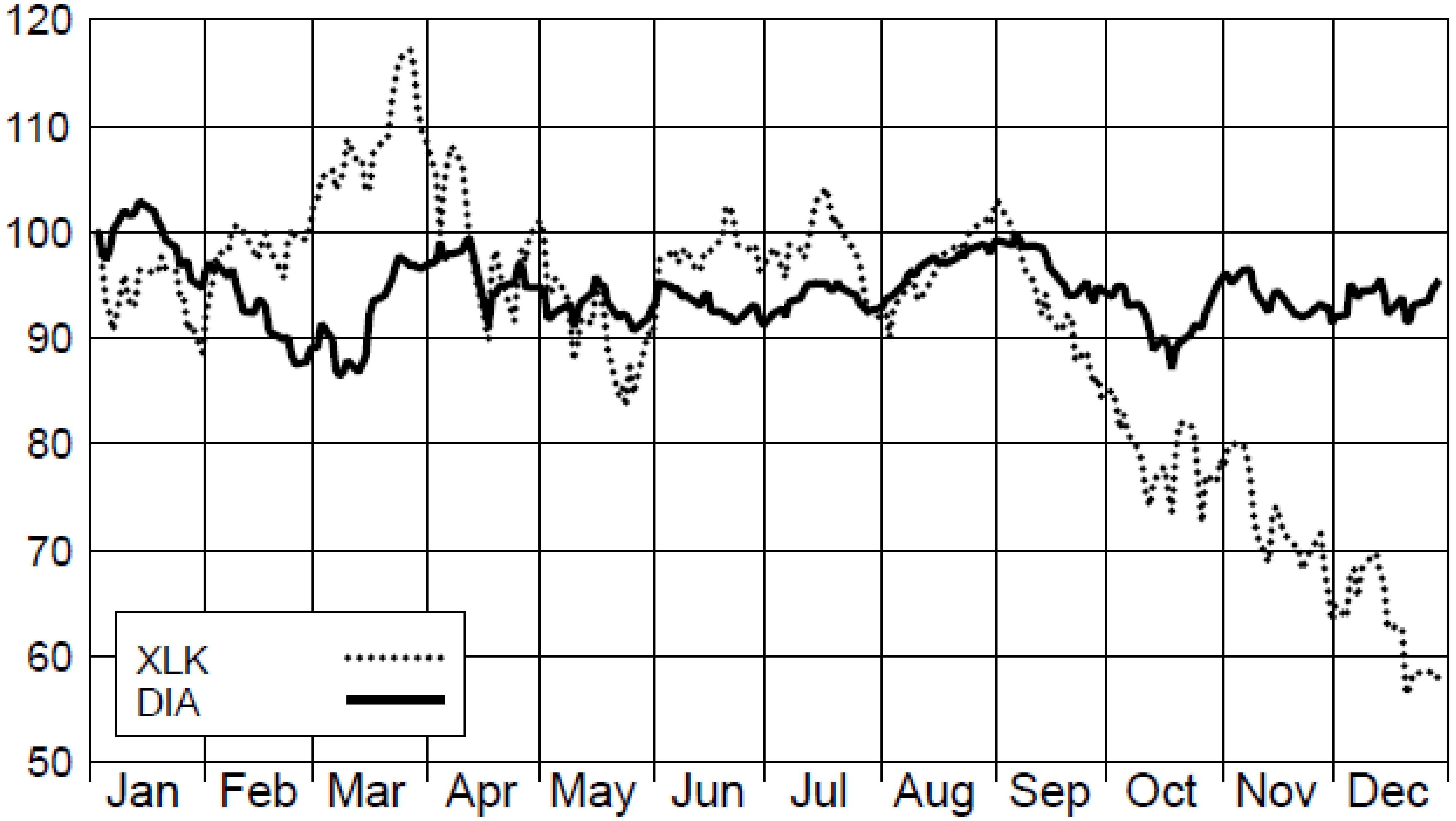
[e] Average of the daily volatility measure which has been calculated as 1,000 times the standard deviation of the MA(1)-demeaned log price changes between 15-minute intervals.

[f] RB and YF indicate Raging Bull and Yahoo! Finance.

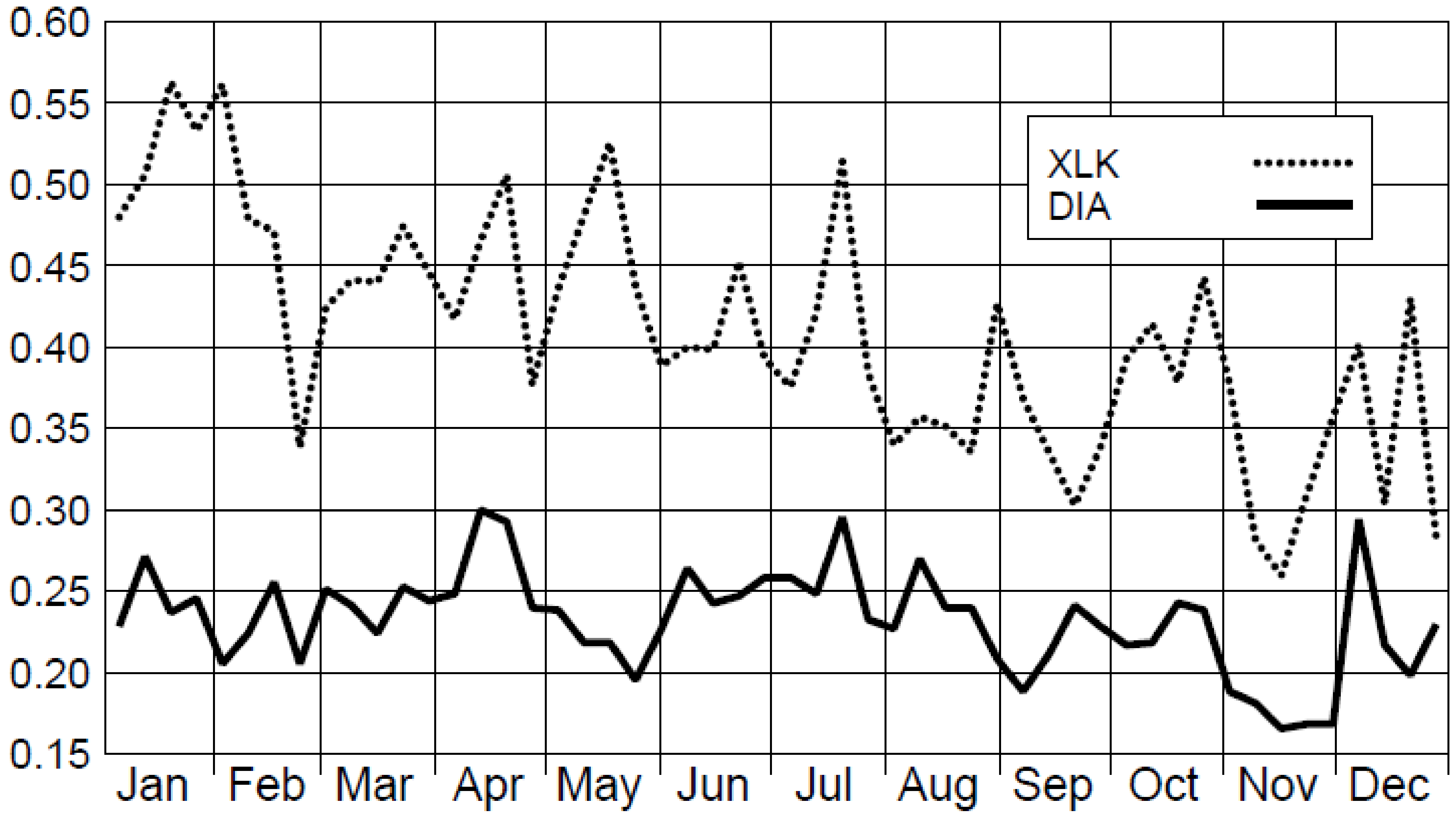
Antweiler and Frank (2004) – Descriptives (3)

Stock index performance and bullishness – AF, Figures 6 + 7

Performance of Dow 30 and telecom index



Bullishness of Dow 30 and telecom index

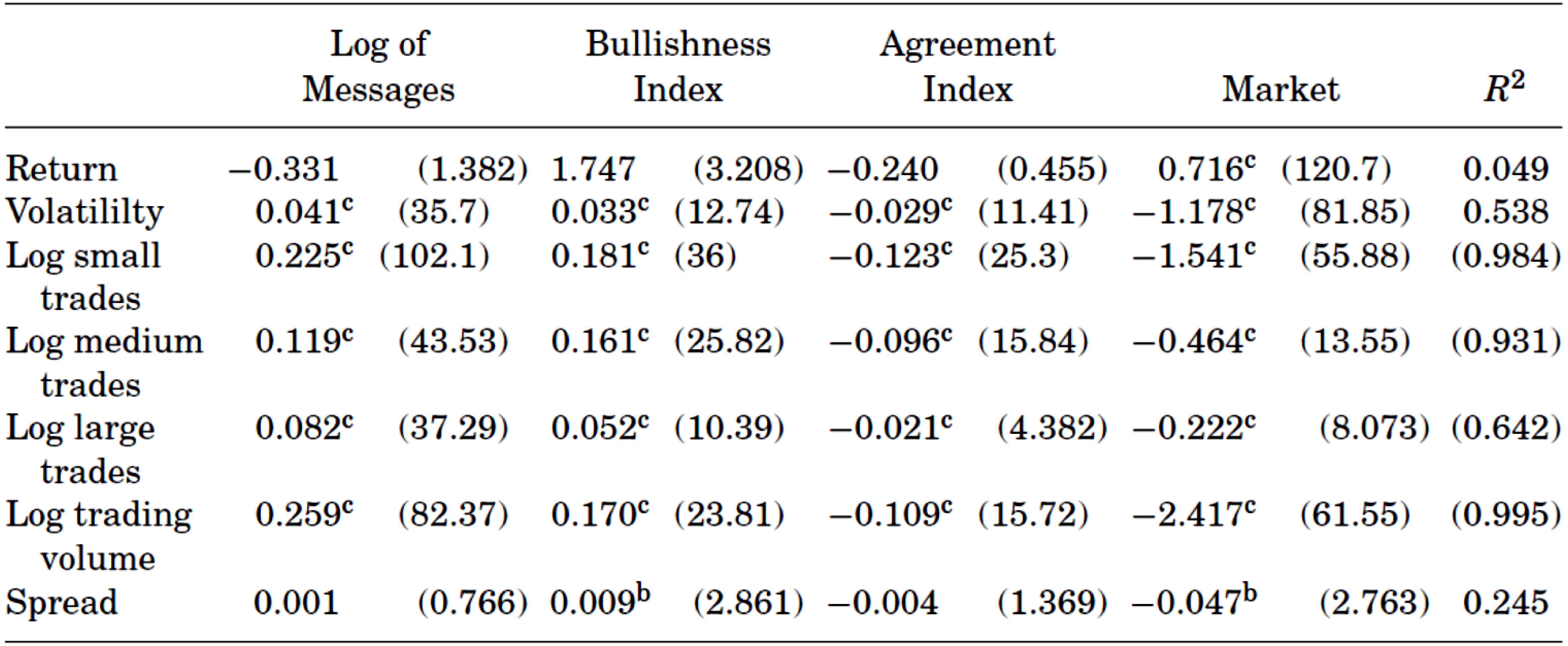


- Figures highlight message posters' biased views of tech companies.
- If message posters are representative for the average (retail) investor these figures nicely illustrate investors' euphoria.

Antweiler and Frank (2004) – Contemporaneous Regressions

Message characteristics and stock market outcomes

- Contemporaneous regressions with 15 minute time intervals – AF, Table 4



	Log of Messages		Bullishness Index		Agreement Index		Market		R^2
Return	−0.331	(1.382)	1.747	(3.208)	−0.240	(0.455)	0.716^c	(120.7)	0.049
Volatililty	0.041^c	(35.7)	0.033^c	(12.74)	-0.029^c	(11.41)	-1.178^c	(81.85)	0.538
Log small trades	0.225^c	(102.1)	0.181^c	(36)	-0.123^c	(25.3)	-1.541^c	(55.88)	(0.984)
Log medium trades	0.119^c	(43.53)	0.161^c	(25.82)	-0.096^c	(15.84)	-0.464^c	(13.55)	(0.931)
Log large trades	0.082^c	(37.29)	0.052^c	(10.39)	-0.021^c	(4.382)	-0.222^c	(8.073)	(0.642)
Log trading volume	0.259^c	(82.37)	0.170^c	(23.81)	-0.109^c	(15.72)	-2.417^c	(61.55)	(0.995)
Spread	0.001	(0.766)	0.009^b	(2.861)	−0.004	(1.369)	-0.047^b	(2.763)	0.245

- Number of messages positively related to volatility and volume.
- More bullishness is associated with high returns, high volatility, and high volume.
- Agreement negatively related to volatility and volume.

Regression setup

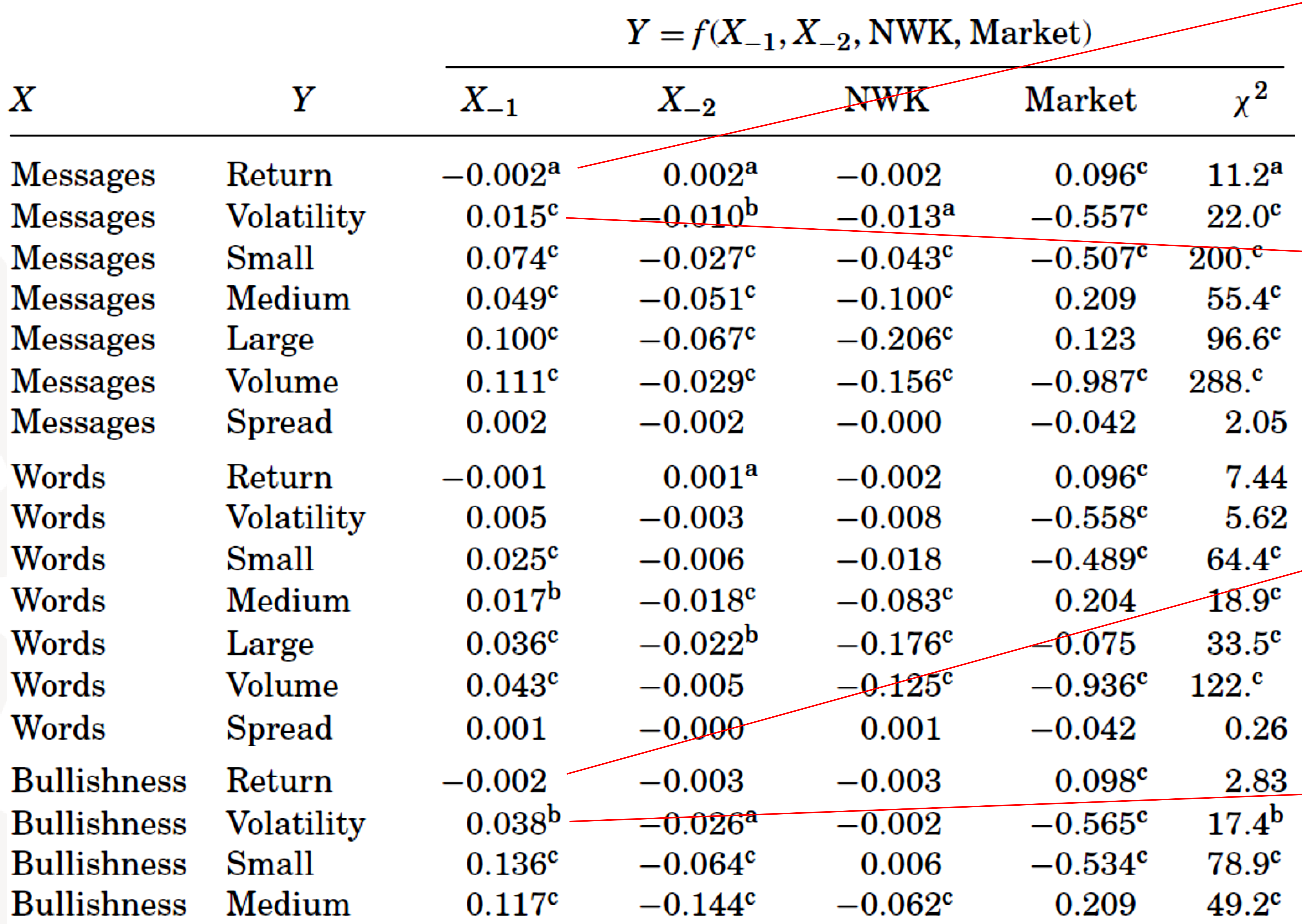
- Each line is a separate regression.
- Numbers in parentheses are t-statistics.
- a, b, and c indicate significance at the 95, 99, and 99.9 percent significance level, respectively.

Antweiler and Frank (2004) – Predictive Regressions

Information content of message characteristics

Predictive regressions at daily frequency – AF, Table 5

- based on Yahoo! Finance messages only



		$Y = f(X_{-1}, X_{-2}, \text{NWK}, \text{Market})$				
X	Y	X_{-1}	X_{-2}	NWK	Market	χ^2
Messages	Return	-0.002^a	0.002^a	-0.002	0.096^c	11.2^a
Messages	Volatility	0.015^c	-0.010^b	-0.013^a	-0.557^c	22.0^c
Messages	Small	0.074^c	-0.027^c	-0.043^c	-0.507^c	$200.^c$
Messages	Medium	0.049^c	-0.051^c	-0.100^c	0.209	55.4^c
Messages	Large	0.100^c	-0.067^c	-0.206^c	0.123	96.6^c
Messages	Volume	0.111^c	-0.029^c	-0.156^c	-0.987^c	$288.^c$
Messages	Spread	0.002	-0.002	-0.000	-0.042	2.05
Words	Return	-0.001	0.001^a	-0.002	0.096^c	7.44
Words	Volatility	0.005	-0.003	-0.008	-0.558^c	5.62
Words	Small	0.025^c	-0.006	-0.018	-0.489^c	64.4^c
Words	Medium	0.017^b	-0.018^c	-0.083^c	0.204	18.9^c
Words	Large	0.036^c	-0.022^b	-0.176^c	-0.075	33.5^c
Words	Volume	0.043^c	-0.005	-0.125^c	-0.936^c	$122.^c$
Words	Spread	0.001	-0.000	0.001	-0.042	0.26
Bullishness	Return	-0.002	-0.003	-0.003	0.098^c	2.83
Bullishness	Volatility	0.038^b	-0.026^a	-0.002	-0.565^c	17.4^b
Bullishness	Small	0.136^c	-0.064^c	0.006	-0.534^c	78.9^c
Bullishness	Medium	0.117^c	-0.144^c	-0.062^c	0.209	49.2^c

Doubling the number of messages → 20 basis points; complete reversal

More messages predict high volatility

Bullishness not statistically significant

More bullishness predicts high volatility

Antweiler and Frank (2004) - Conclusion

Summary and conclusions

- Results of contemporaneous regressions are economically plausible.
- Informational content of messages seems questionable.
  - Economically small positive relation to future returns.
  - Returns are followed by reversal.
  - # of messages and optimism predict higher volatility.

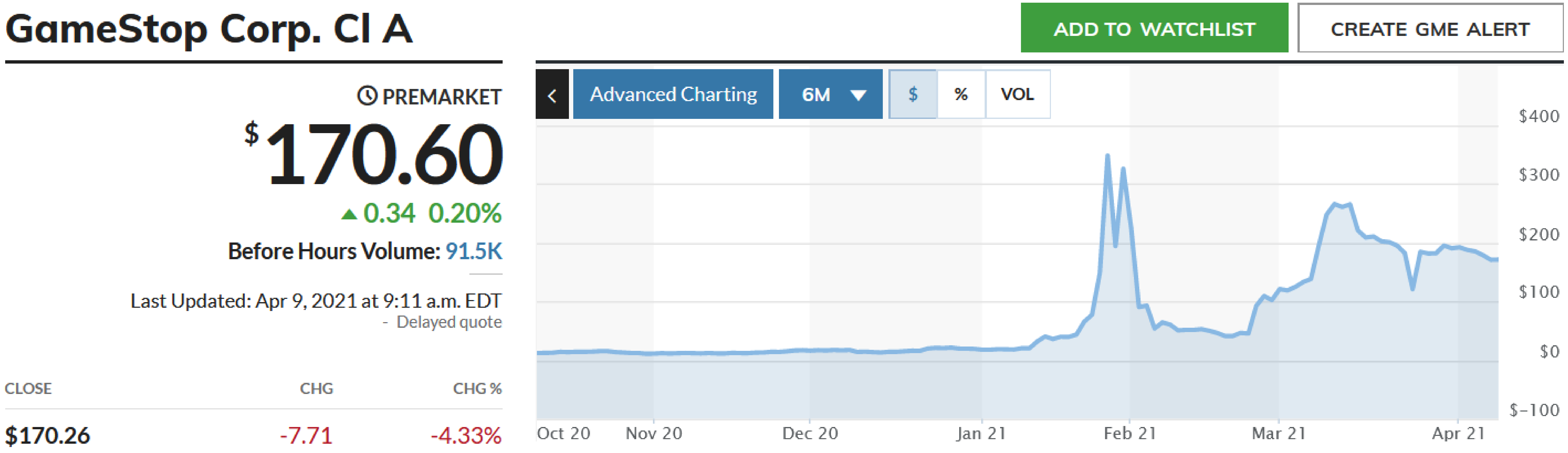
  → Stock message board activity and tone may indicate investor sentiment and proxy for noise trading activity.

We will discuss their methodology later when we know more about machine learning!

Reddit WallStreetBets (1)

WallStreetBets on Reddit

- Reddit's WallStreetBets is reported to be a key driver of the Gamestop stock frenzy.



GameStop Corp. Cl A

PREMARKET

$170.60

▲0.34 0.20%

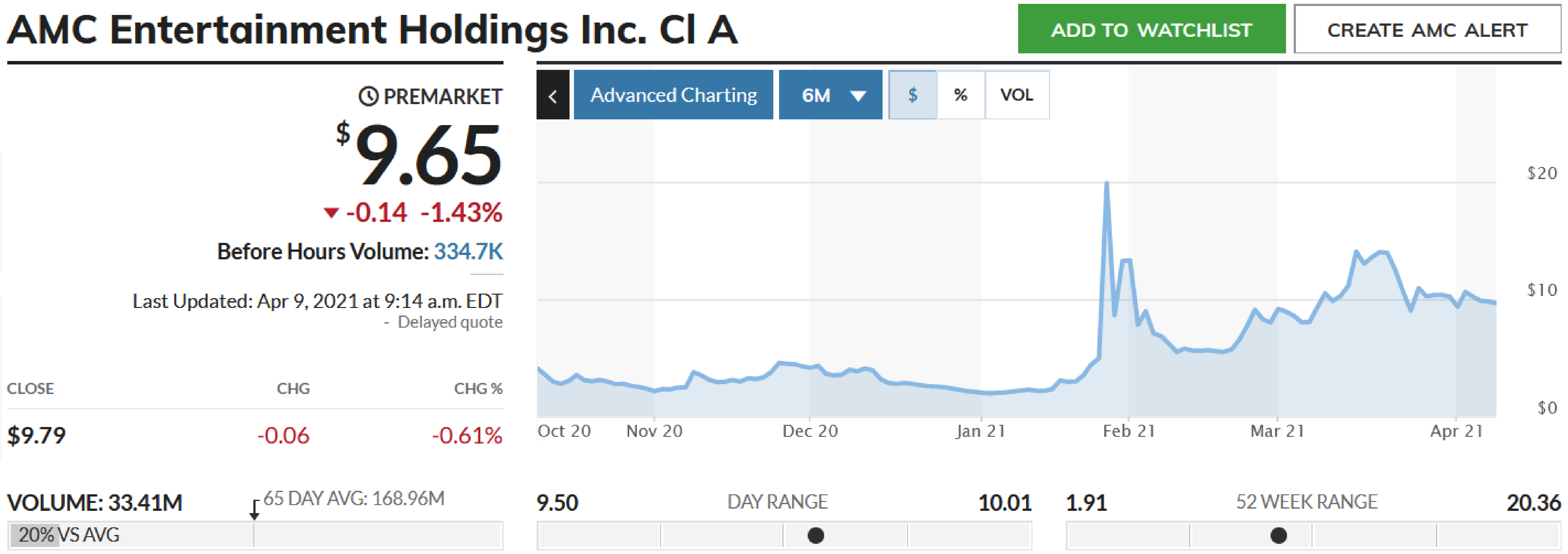
Before Hours Volume: 91.5K

Last Updated: Apr 9, 2021 at 9:11 a.m. EDT
- Delayed quote

CLOSE	CHG	CHG %
$170.26	-7.71	-4.33%

WallStreetBets on Reddit

- Other stocks also affected by the frenzy.



AMC Entertainment Holdings Inc. Cl A

PREMARKET

$9.65

▼ -0.14 -1.43%

Before Hours Volume: 334.7K

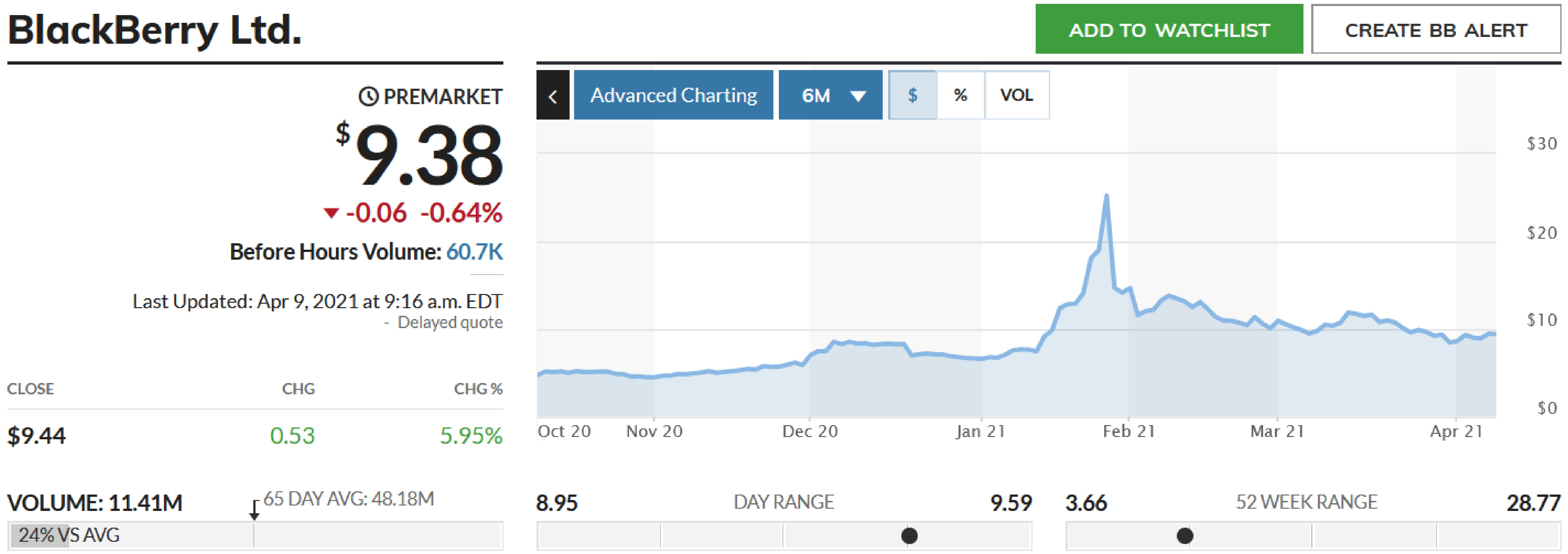
Last Updated: Apr 9, 2021 at 9:14 a.m. EDT
- Delayed quote

CLOSE	CHG	CHG %
$9.79	-0.06	-0.61%

VOLUME: 33.41M 65 DAY AVG: 168.96M
20% VS AVG

9.50 DAY RANGE 10.01

1.91 52 WEEK RANGE 20.36



BlackBerry Ltd.

PREMARKET

$9.38

▼ -0.06 -0.64%

Before Hours Volume: 60.7K

Last Updated: Apr 9, 2021 at 9:16 a.m. EDT
- Delayed quote

CLOSE	CHG	CHG %
$9.44	0.53	5.95%

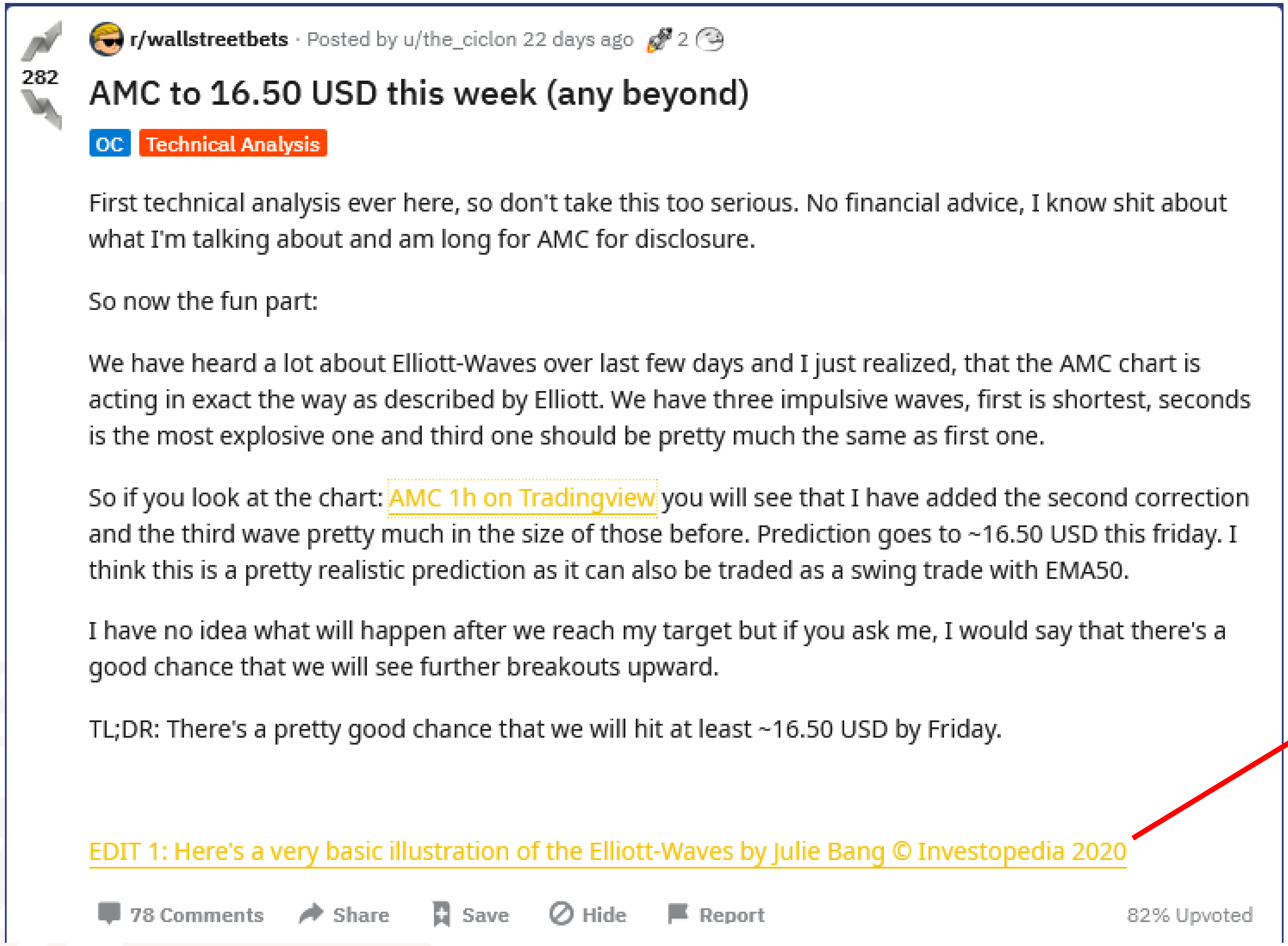
VOLUME: 11.41M 65 DAY AVG: 48.18M
24% VS AVG

8.95 DAY RANGE 9.59

3.66 52 WEEK RANGE 28.77

WallStreetBets on Reddit – Example: Post on AMC

Source:
https://new.reddit.com/r/wallstreetbets/comments/m75itq/amc_to_1650_usd_this_week_any_beyond/



r/wallstreetbets · Posted by u/the_ciclon 22 days ago 2

282

AMC to 16.50 USD this week (any beyond)

OC Technical Analysis

First technical analysis ever here, so don't take this too serious. No financial advice, I know shit about what I'm talking about and am long for AMC for disclosure.

So now the fun part:

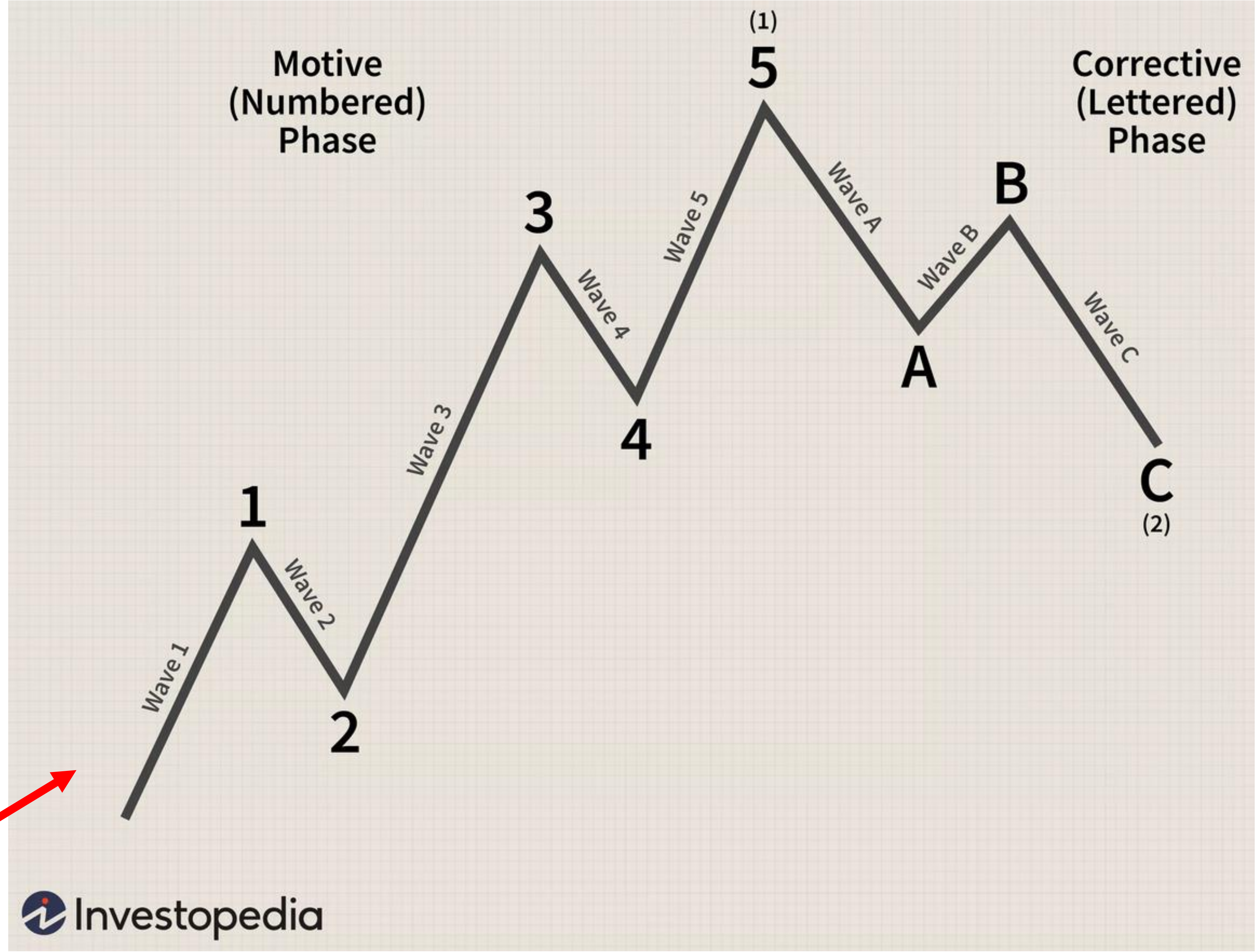
We have heard a lot about Elliott-Waves over last few days and I just realized, that the AMC chart is acting in exact the way as described by Elliott. We have three impulsive waves, first is shortest, seconds is the most explosive one and third one should be pretty much the same as first one.

So if you look at the chart: AMC 1h on Tradingview you will see that I have added the second correction and the third wave pretty much in the size of those before. Prediction goes to ~16.50 USD this friday. I think this is a pretty realistic prediction as it can also be traded as a swing trade with EMA50.

I have no idea what will happen after we reach my target but if you ask me, I would say that there's a good chance that we will see further breakouts upward.

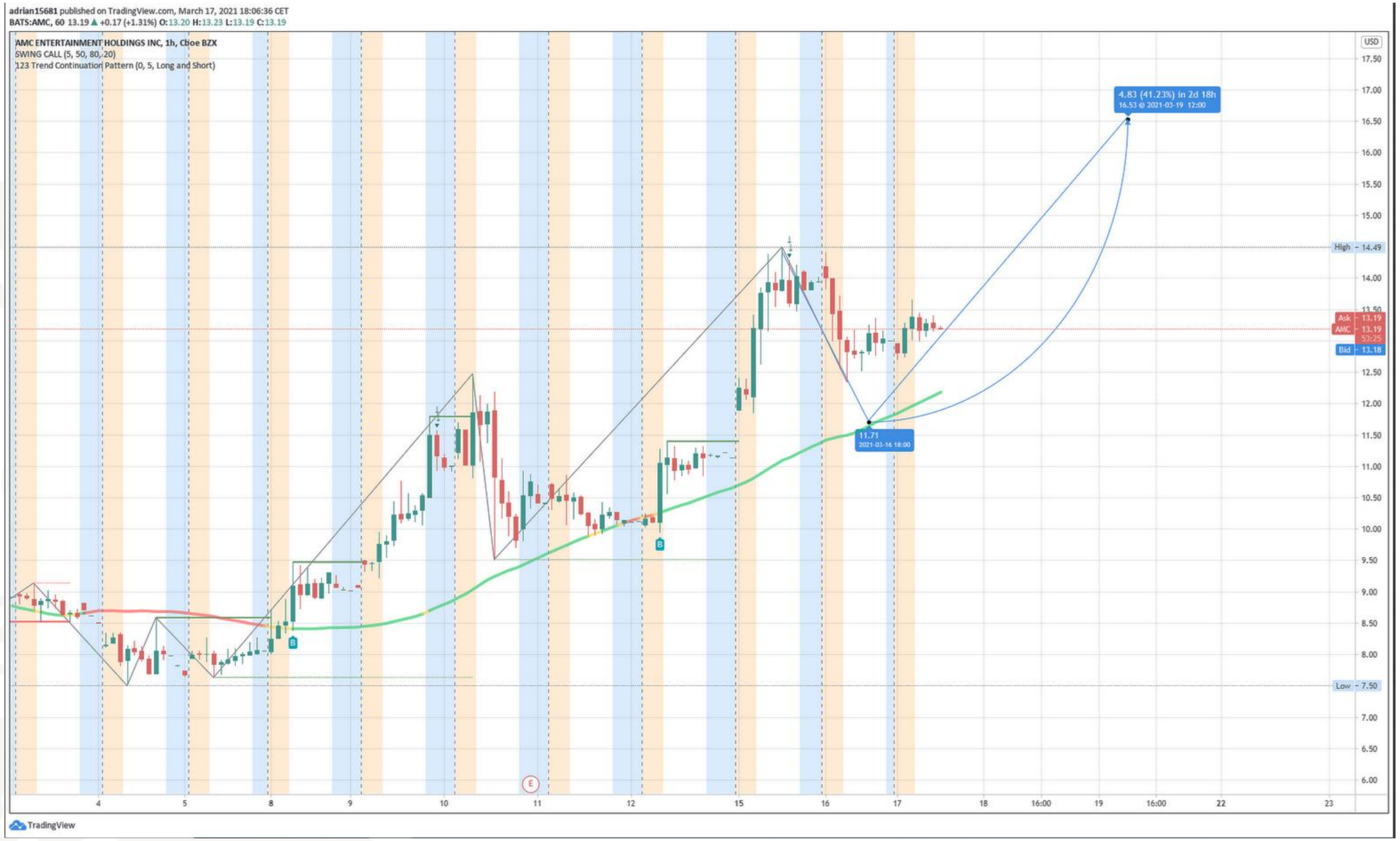
TL;DR: There's a pretty good chance that we will hit at least ~16.50 USD by Friday.

EDIT 1: Here's a very basic illustration of the Elliott-Waves by Julie Bang © Investopedia 2020



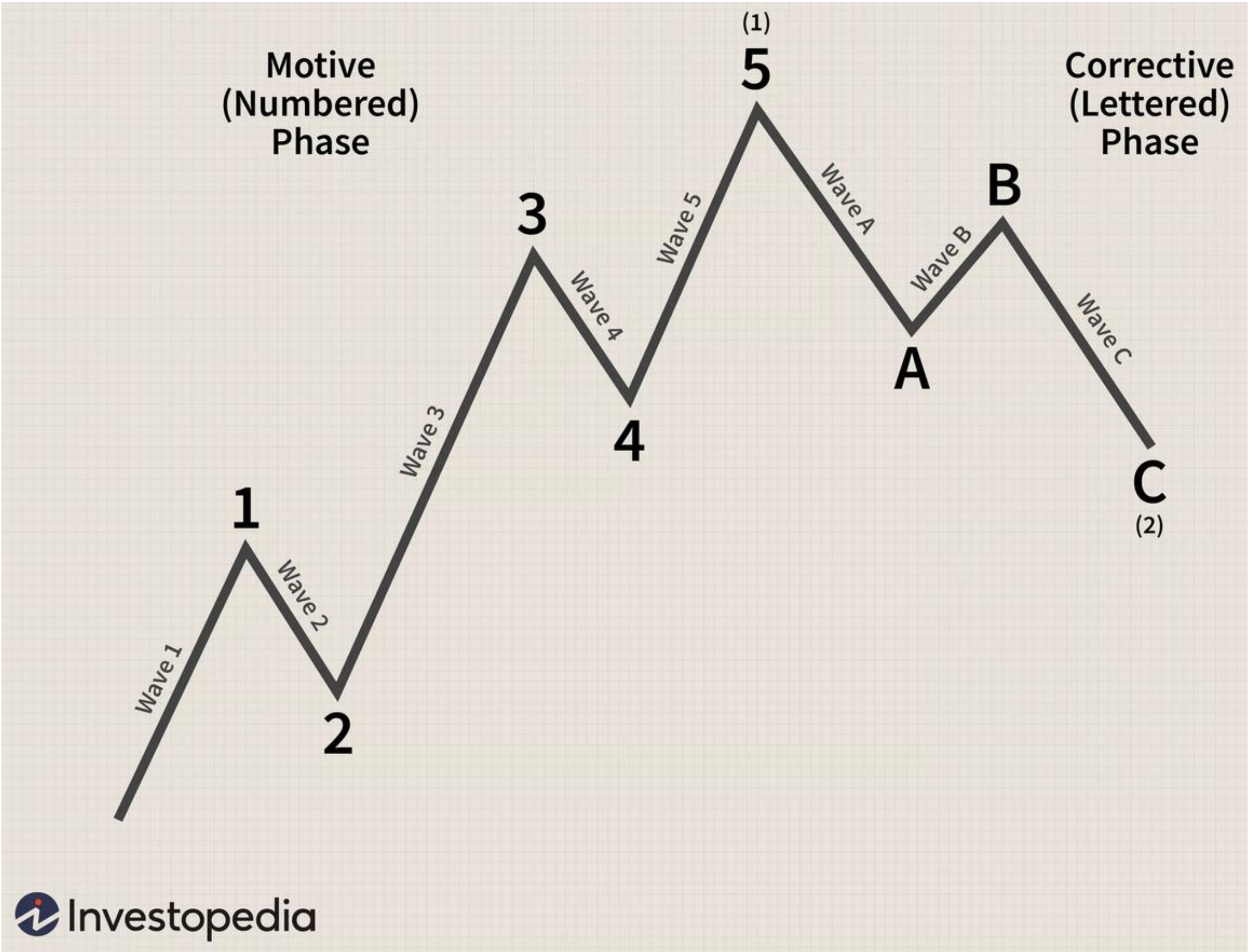
Intro to Elliott Wave Theory
https://www.investopedia.com/articles/technical/111401.asp

WallStreetBets on Reddit – Example: Post on AMC



adrian15681 published on TradingView.com, March 17, 2021 18:06:36 CET
BATS:AMC, 60 13.19 ▲ +0.17 (+1.31%) O:13.20 H:13.23 L:13.19 C:13.19

AMC ENTERTAINMENT HOLDINGS INC, 1h, Cboe BZX
SWING CALL (5, 50, 80, 20)
123 Trend Continuation Pattern (0, 5, Long and Short)



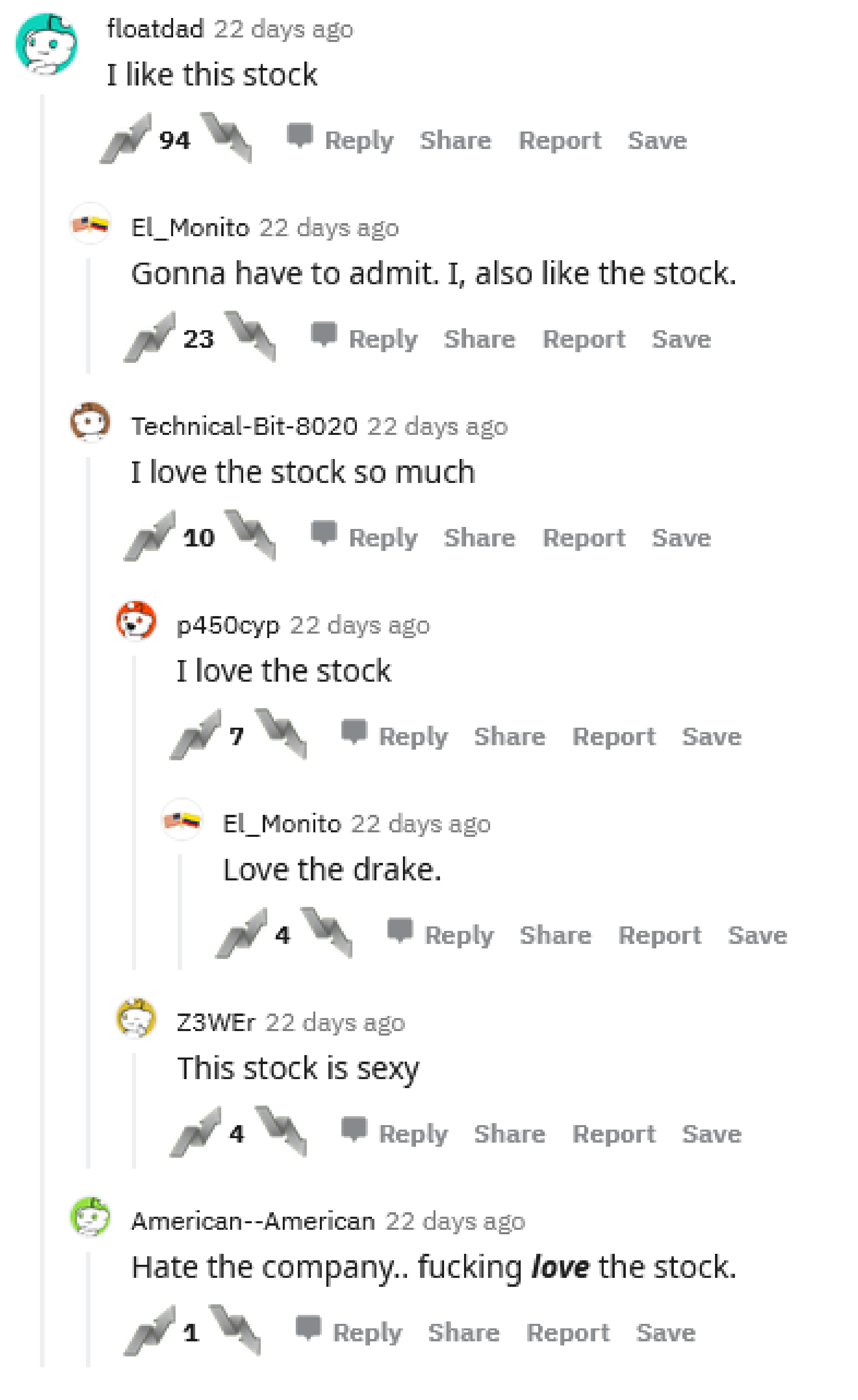
Source: https://new.reddit.com/r/wallstreetbets/comments/m75itq/amc_to_1650_usd_this_week_any_beyond/

Reddit WallStreetBets (5)

WallStreetBets on Reddit – Example: Post on AMC

- Some comments on this “technical analysis”

Source:
https://new.reddit.com/r/wallstreetbets/comments/m75itq/amc_to_1650_usd_this_week_any_beyond/



floatdad 22 days ago
I like this stock
94 Reply Share Report Save

El_Monito 22 days ago
Gonna have to admit. I, also like the stock.
23 Reply Share Report Save

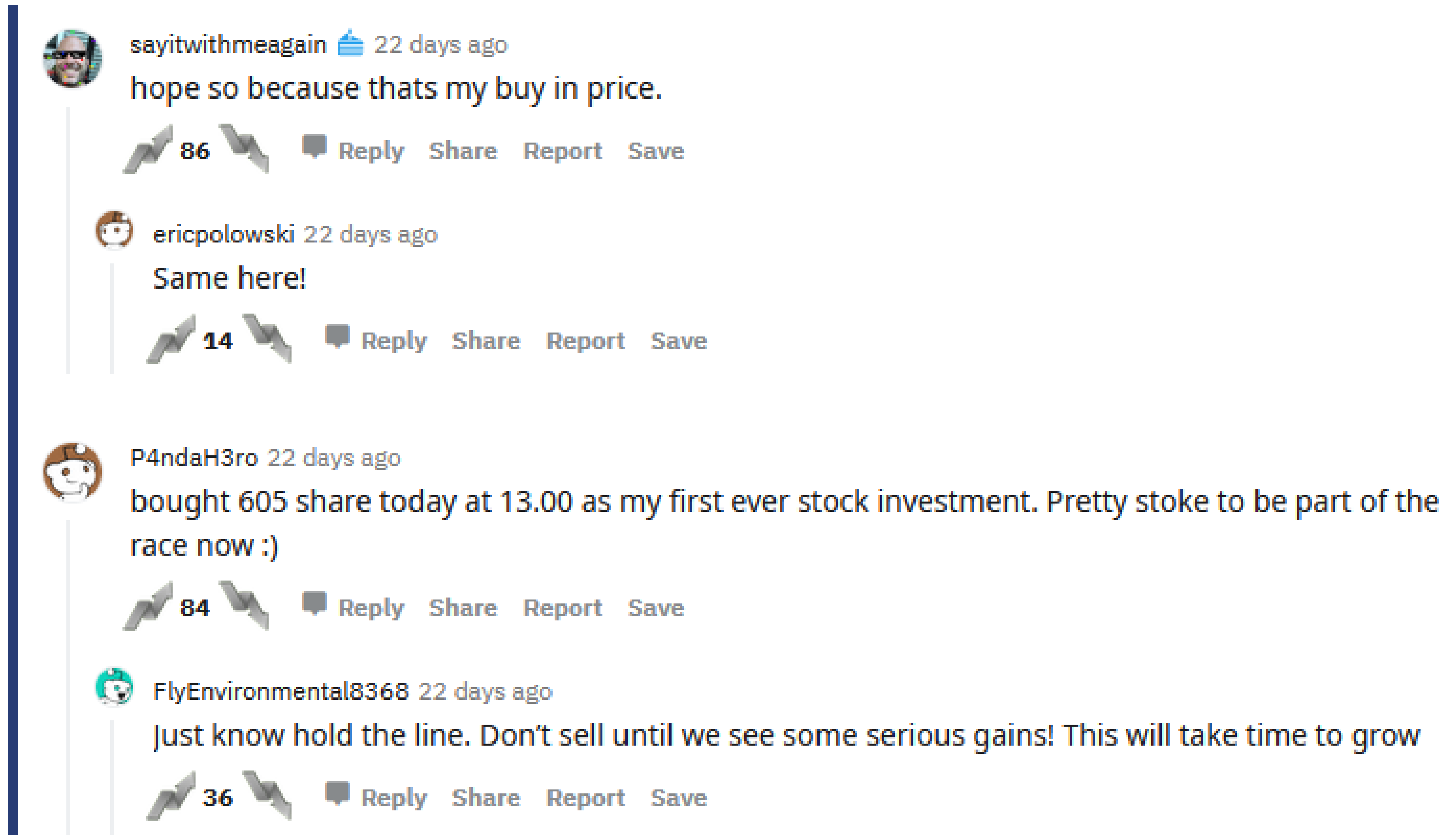
Technical-Bit-8020 22 days ago
I love the stock so much
10 Reply Share Report Save

p450cyp 22 days ago
I love the stock
7 Reply Share Report Save

El_Monito 22 days ago
Love the drake.
4 Reply Share Report Save

Z3WEr 22 days ago
This stock is sexy
4 Reply Share Report Save

American--American 22 days ago
Hate the company.. fucking *love* the stock.
1 Reply Share Report Save



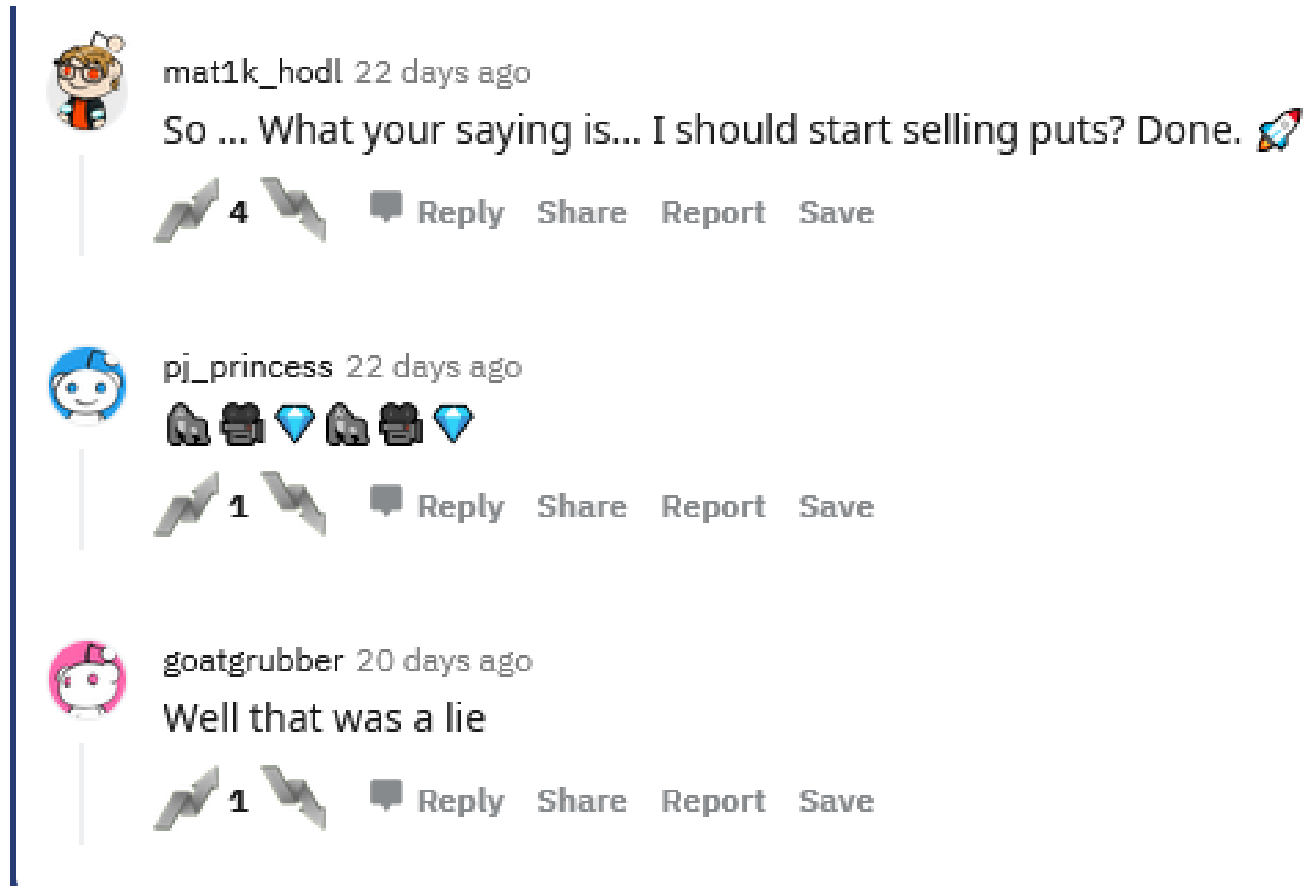
sayitwithmeagain 22 days ago
hope so because thats my buy in price.
86 Reply Share Report Save

ericpolowski 22 days ago
Same here!
14 Reply Share Report Save

P4ndaH3ro 22 days ago
bought 605 share today at 13.00 as my first ever stock investment. Pretty stoke to be part of the race now :)
84 Reply Share Report Save

FlyEnvironmental8368 22 days ago
Just know hold the line. Don't sell until we see some serious gains! This will take time to grow
36 Reply Share Report Save

Last (most recent) three comments on the analysis



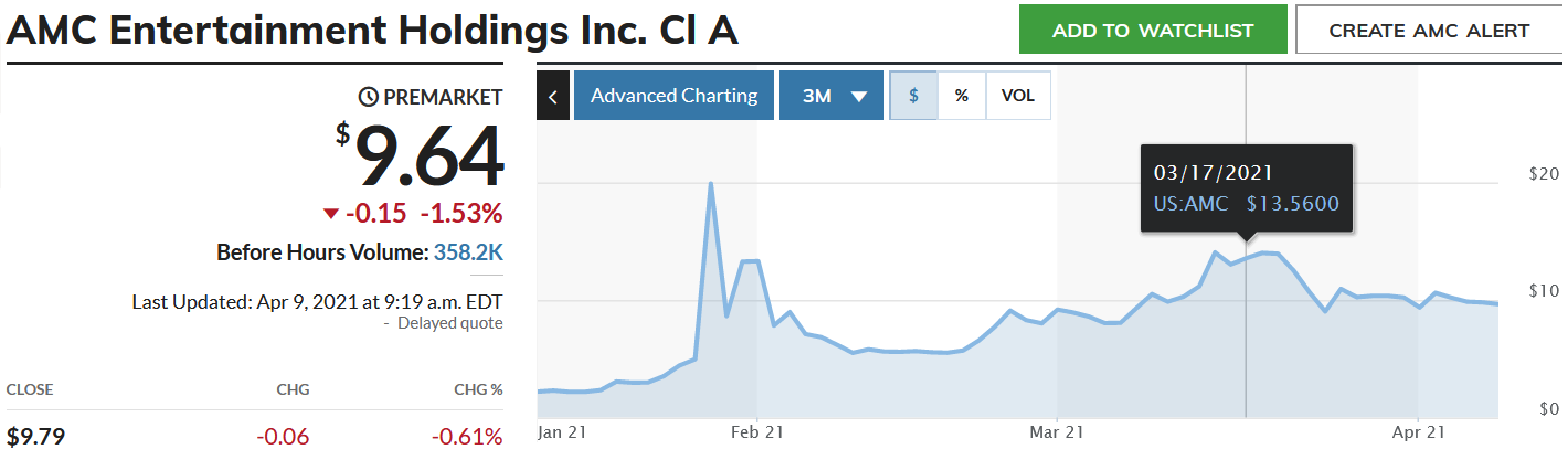
mat1k_hodl 22 days ago
So ... What your saying is... I should start selling puts? Done. 🚀
4 Reply Share Report Save

pj_princess 22 days ago
🦍✊💎🦍✊💎
1 Reply Share Report Save

goatgrubber 20 days ago
Well that was a lie
1 Reply Share Report Save

WallStreetBets on Reddit – Example: Post on AMC

- Did the recommendation work?
- Post was on Wednesday March 17 → look at price on Friday March 19.



AMC Entertainment Holdings Inc. Cl A

PREMARKET

$9.64

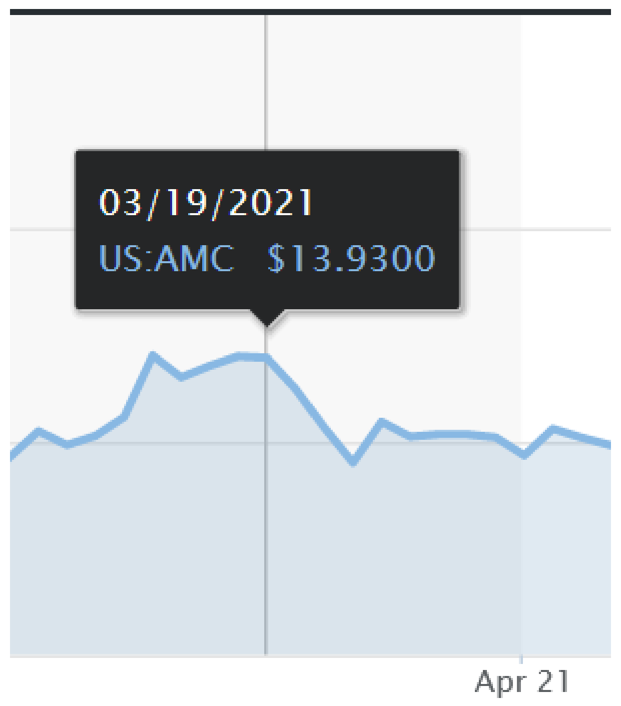
▼ -0.15 -1.53%

Before Hours Volume: 358.2K

Last Updated: Apr 9, 2021 at 9:19 a.m. EDT
- Delayed quote

CLOSE	CHG	CHG %
$9.79	-0.06	-0.61%

03/17/2021
US:AMC $13.5600



03/19/2021
US:AMC $13.9300

WallStreetBets on Reddit – Conclusion

- Might be today's equivalent to the Yahoo! Finance and Raging Bull discussion boards.
- However, people mostly discuss Gamestop and a few other popular stocks like AMC.
- While it might be interesting to quantify people's sentiment, I see little potential for top research on WallStreetBets (evidence based on N ~ 1).

- More promising avenue for research: understanding how investors select information.
  Cookson, Engelberg, and Mullins (2022, WP): "Echo Chambers"
  https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3603107
  Abstract:
  "*We find evidence of selective exposure to confirmatory information among 400,000 users on the investor social network StockTwits. Self-described bulls are 5 times more likely to follow a user with a bullish view of the same stock than self-described bears. Consequently, bulls see 62 more bullish messages and 24 fewer bearish messages than bears over the same 50-day period. These "echo chambers" exist even among professional investors and are strongest for investors who trade on their beliefs. Finally, beliefs formed in echo chambers are associated with lower ex-post returns, more siloing of information and more trading volume.*"

Introduction to Machine learning

With a focus on text data

Machine Learning (1)

Machine learning with text data

- Antweiler and Frank (2004) want to predict the sentiment of a message from its text.
- This prediction task is ubiquitous in business and economics, e.g., customer reviews, newspaper articles, tweets, central bank speeches.
- Text can be represented as a vector of words.
  - Text 1: “Sales strongly increased last year.”
  - Text 2: “Sales strongly declined last quarter.”
  - Vocabulary → word vector: $(sales, strongly, increased, last, year, declined, quarter)$
  - $x_{Text\ 1} = (1,1,1,1,1,0,0)$
  - $x_{Text\ 2} = (1,1,0,1,0,1,1)$
- Outcome variable: Sentiment $y = 1$ (positive sentiment) or $y = 0$ (negative sentiment)
  - Example: $y_1 = 1$, y_2 =0

Machine learning with text data – continued

- Train an algorithm to get from x (text) to y (variable of interest, e.g., sentiment)

$$\begin{tikzcd} \text{Training data } (y, x) \arrow[r] & \text{Application data } (\hat{y} = \hat{f}(x), x) \end{tikzcd}$$

- Goal: small $E\left(y - \hat{f}(x)\right)^2$
- Start with a linear regression: $\hat{f}(x) = \hat{\beta}'x = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$
- From training data, pick $\hat{\beta}$ with best in-sample fit:

$$\min_{\hat{\beta}} E\left(y - \hat{\beta}'x\right)^2 \rightarrow \min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}'x_i\right)^2$$

- Property: **B**est **L**inear **U**nbiased **E**stimator (Gauss-Markov)
  → optimal for out-of-sample prediction?

Bias – variance decomposition / trade-off

- New data point: $y = \beta' x + \epsilon \;\; (E[\varepsilon|x] = 0)$
- Loss at new data point: $(\hat{y} - y)^2 = (\hat{\beta}' x - \beta' x - \epsilon)^2$

- Average over draws of the training sample:

$$E_{T,\varepsilon}[(\hat{y} - y)^2] = E_T\left[\left(\hat{\beta}' x - \beta' x\right)^2\right] + E_\varepsilon[\varepsilon^2]$$

$$= \underbrace{\left(\left(E_T[\hat{\beta}] - \beta\right)' x\right)^2}_{\text{Bias or Approximation}} + \underbrace{x' Var_T\left(\hat{\beta}\right) x}_{\text{Variance or Overfit}} + \underbrace{Var_\varepsilon(\varepsilon|x)}_{\text{Irreducible noise}}$$

Takeaway
High approximation quality means that algorithms does not only fit the true relation but also random noise → high variance.

Bias: algorithm misses relevant relations between features (x) and outcome (y)

Variance: too high sensitivity to small fluctuations in the training data. Algorithm models random noise.

Overfitting problem

- Overfitting problem does not only apply to linear regressions.
- For more complex methods, it is even more severe:
  More flexible methods (non-linearities, interaction) → better approximation quality but also higher overfit.
- Text data are ultra high dimensional!
  - Often more variables than observations → k (# of coefficients) typically larger than n (# of obs.).
    In our example: $n = 2$ and $\mathrm{k} = 7$.
  - Very high approximation quality resulting in massive overfit.

Conclusion:

1. Flexible functional form → good approximation quality
2. Limit expressiveness / “regularize” → avoid overfit

Regularization

- Instead of standard OLS: $\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2$
- Fit a constrained problem: $\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2$ s.t. $\|\hat{\beta}\| \leq c$
- Typical norms
  1. Limit # of non-zero coefficients: $\|\hat{\beta}\|_0 = \sum_{\text{j}=1}^{k} 1_{\hat{\beta} \neq 0}$
     → sparse solution but computationally infeasible to run.
  2. Limit sum of abs. coefficient size: $\|\hat{\beta}\|_1 = \sum_{\text{j}=1}^{k} |\hat{\beta}_j|$ → LASSO
  3. Limit sum of squared coefficient size: $\|\hat{\beta}\|_2^2 = \sum_{\text{j}=1}^{k} \hat{\beta}_j^2$ → Ridge
- Intercept ($\hat{\beta}_0$) not penalized.
- Normalize covariates (x).

LASSO regression

- Limit sum of abs. coefficient size.

$$\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2 \text{ s.t. } \sum_{j=1}^{k} \left|\hat{\beta}_j\right| \leq c$$

- In practice, solve the Lagrange relaxation of the problem:

$$\min_{\hat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2 + \lambda \sum_{j=1}^{k} \left|\hat{\beta}_j\right|$$

- Properties of LASSO:
  - Selects and shrinks
  - Produces sparse solutions: many of variables exactly zero.
  - “Capitalist”: in doubt give all to one variable

Machine Learning (7)

Ridge regression

- Limit sum of squared coefficient size

$$\min_{\widehat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2 \text{ s.t. } \sum_{\mathrm{j}=1}^{k} \hat{\beta}_j^2 \ \leq c$$

- In practice, solve the Lagrange relaxation of the problem:

$$\min_{\widehat{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}' x_i\right)^2 + \lambda \sum_{j=1}^{k} \hat{\beta}_j^{\ 2}$$

- <u>Properties of Ridge</u>:
  - Shrinks towards zero but not exactly zero
  - “Socialist”: in doubt distribute to multiple

Choosing regularization parameter

→ Goal: best out-of-sample fit.

Example:

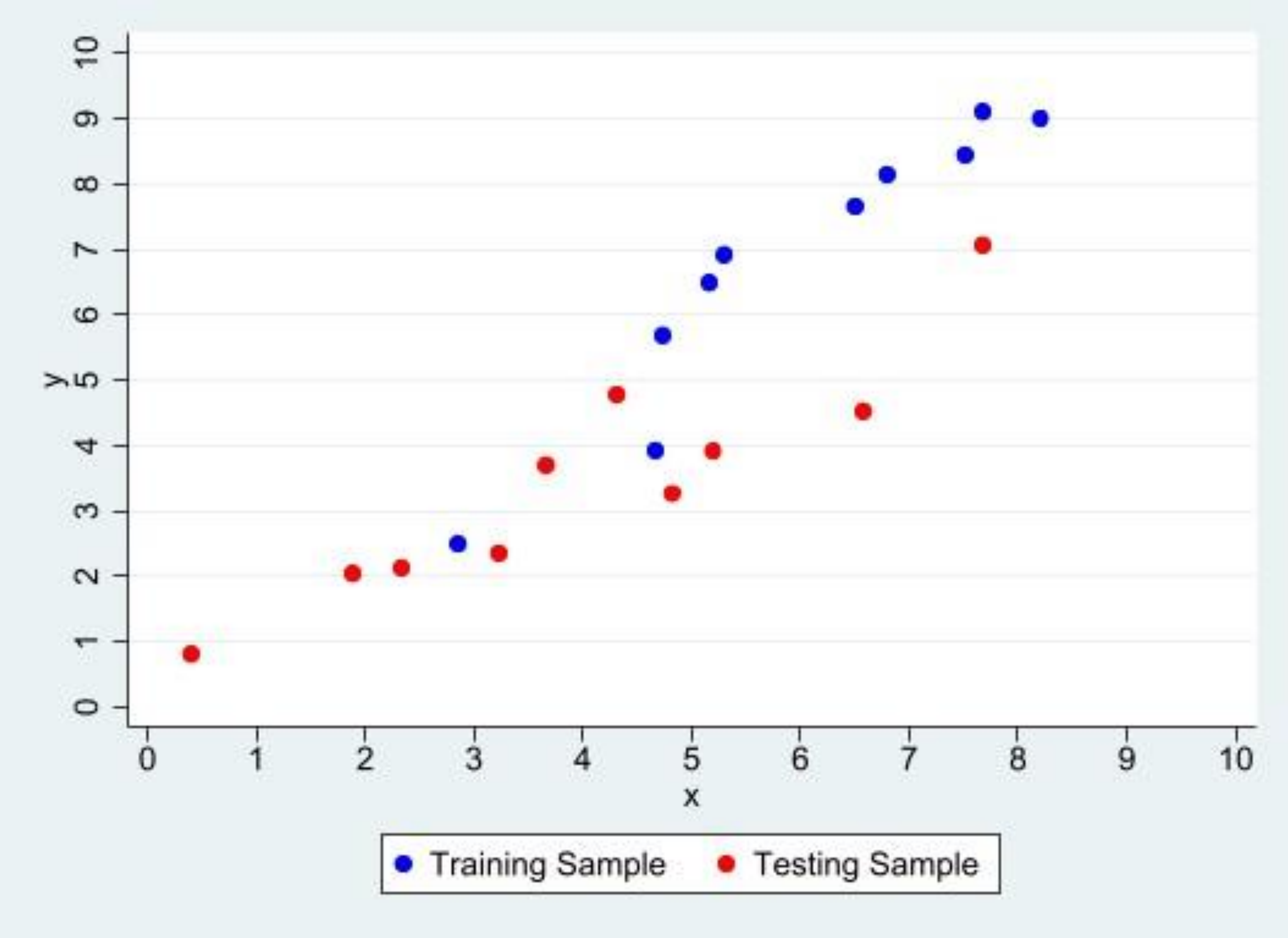
- 20 data points (10 train, 10 test)
- For illustration (2 dimensional), we look at a simple binary choice:
  - Simple model: linear term

    $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

    In LASSO, Ridge → high λ
  - Complex model: linear + quadratic term

    $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2$

    In LASSO, Ridge → low λ

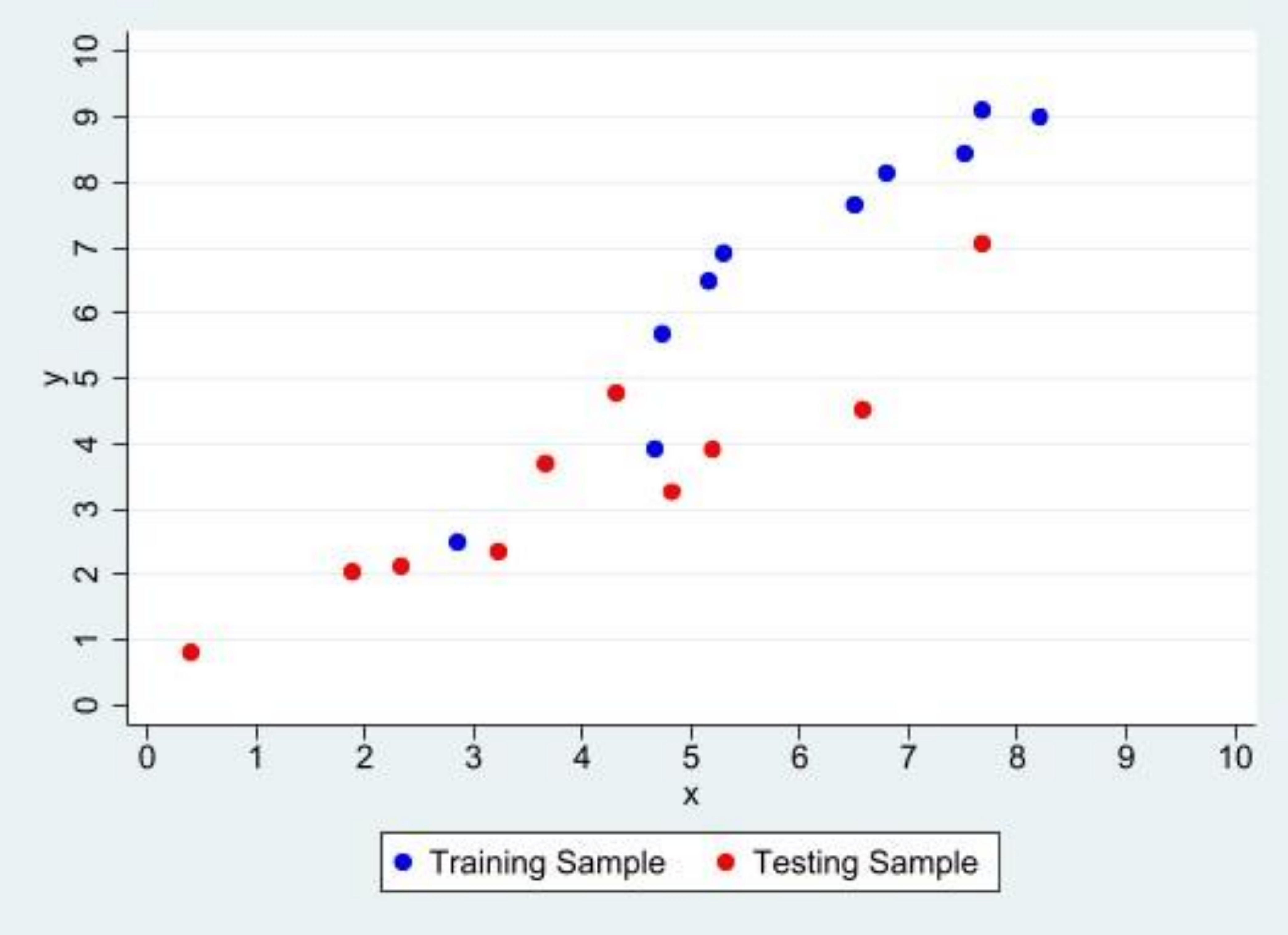


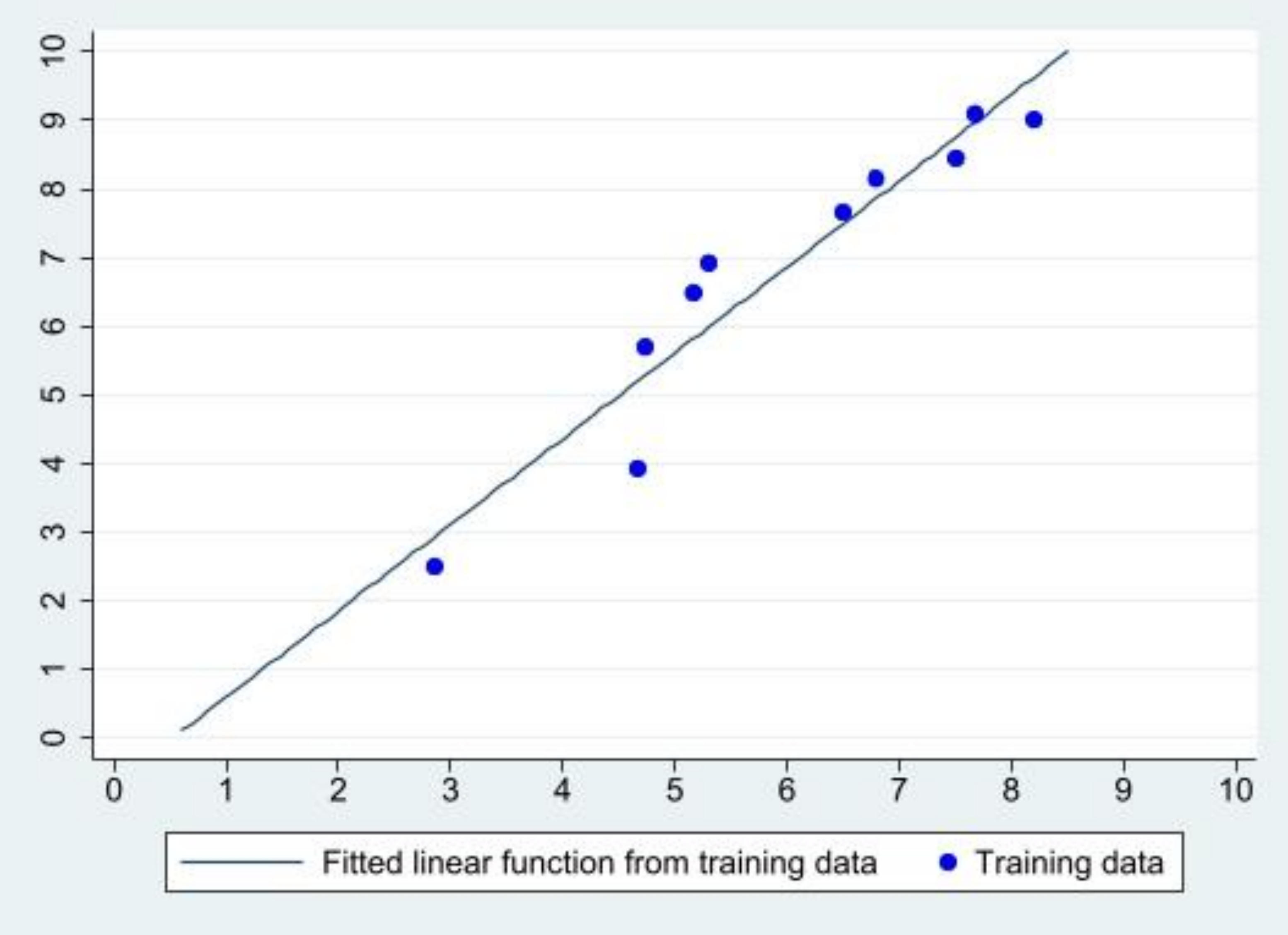
Machine Learning (9)

Simple model (high λ)

→ Goal: best out-of-sample fit.

Example: 20 data points (10 train, 10 test)

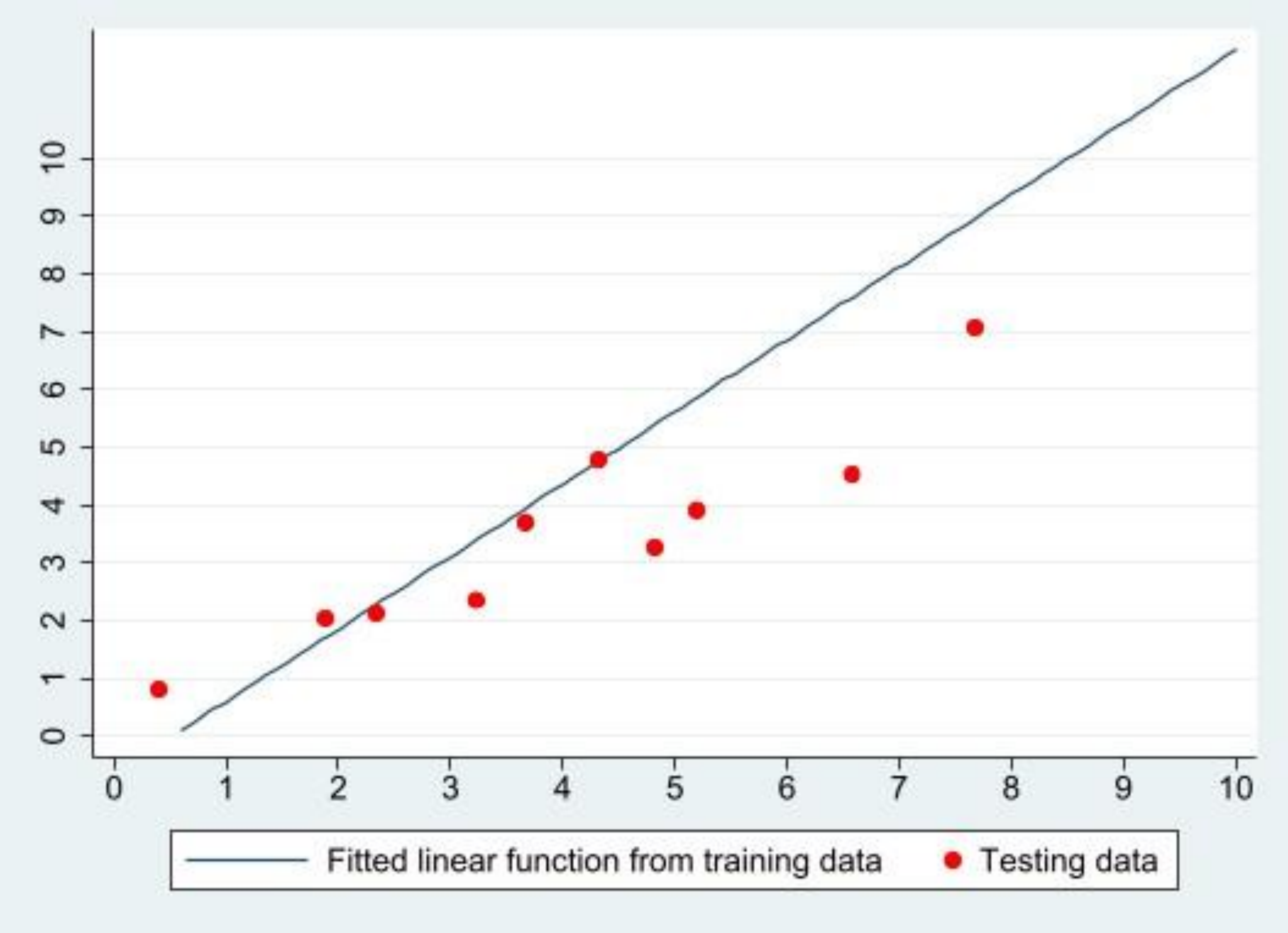




$RMSE = 0.70$

Use training data to fit linear model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

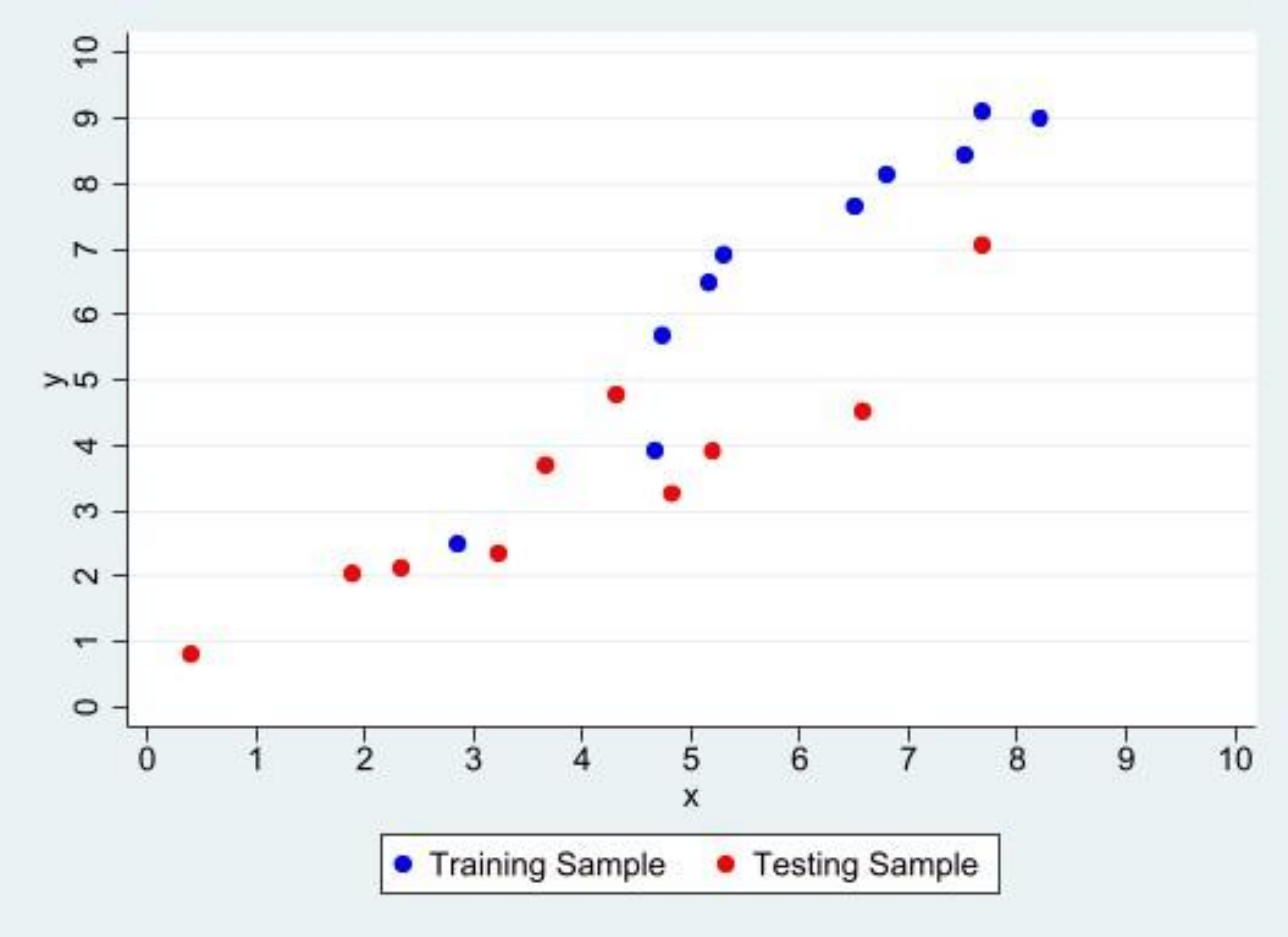


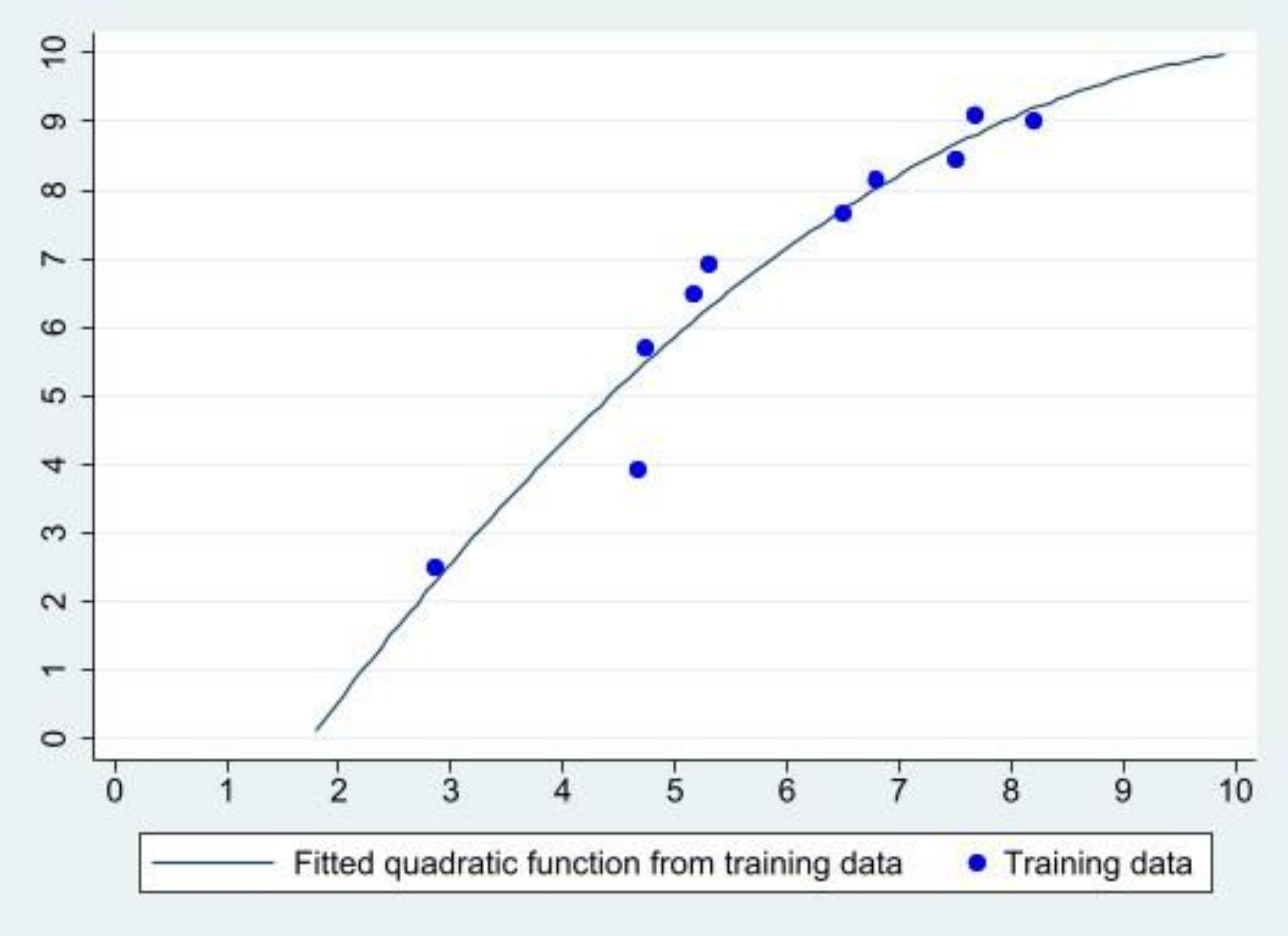
$RMSE = 1.53$

Complex model (low λ)

→ Goal: best out-of-sample fit.

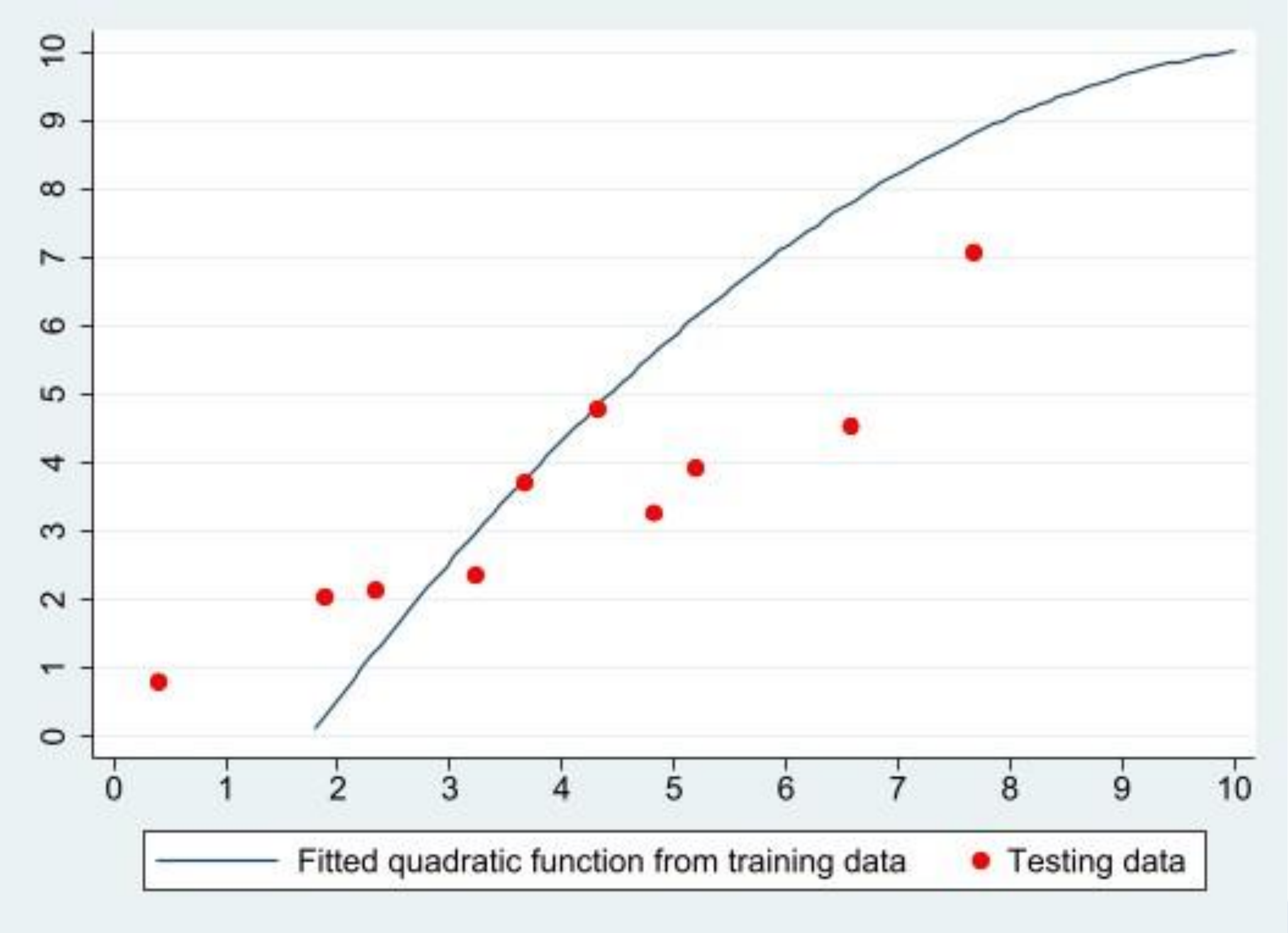
Example: 20 data points (10 train, 10 test)





$RMSE = 0.65$

Use training data to fit quadratic model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2$



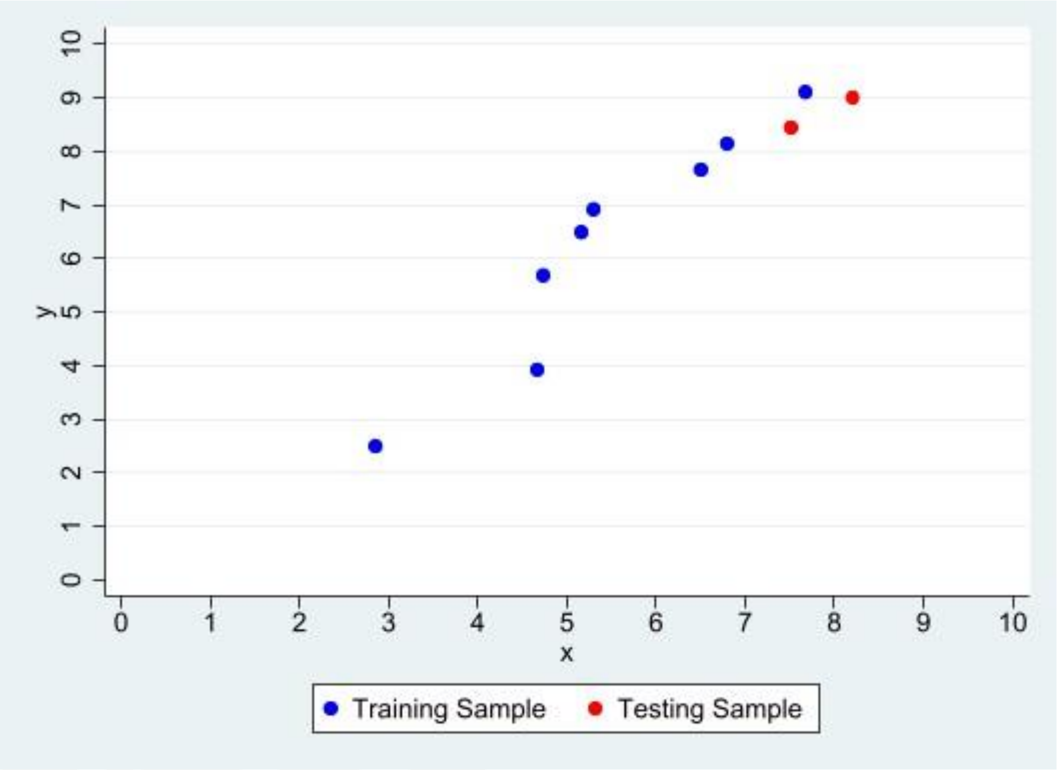
Quadratic model too complex → overfitting.

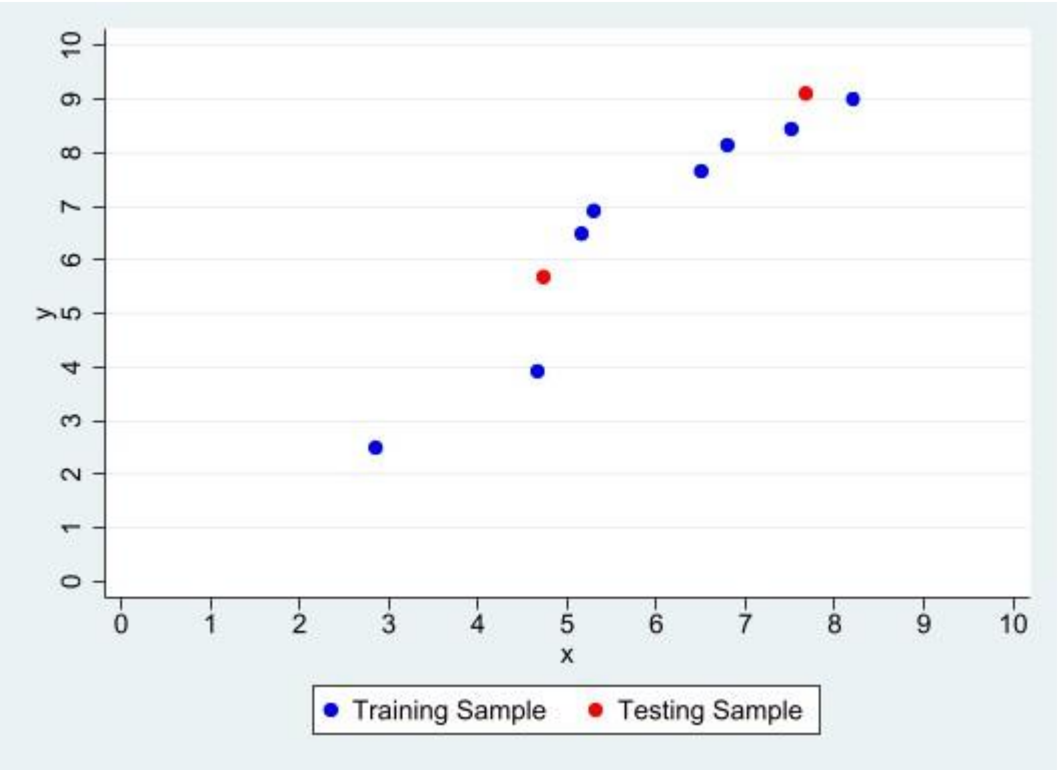
$RMSE = 2.09$

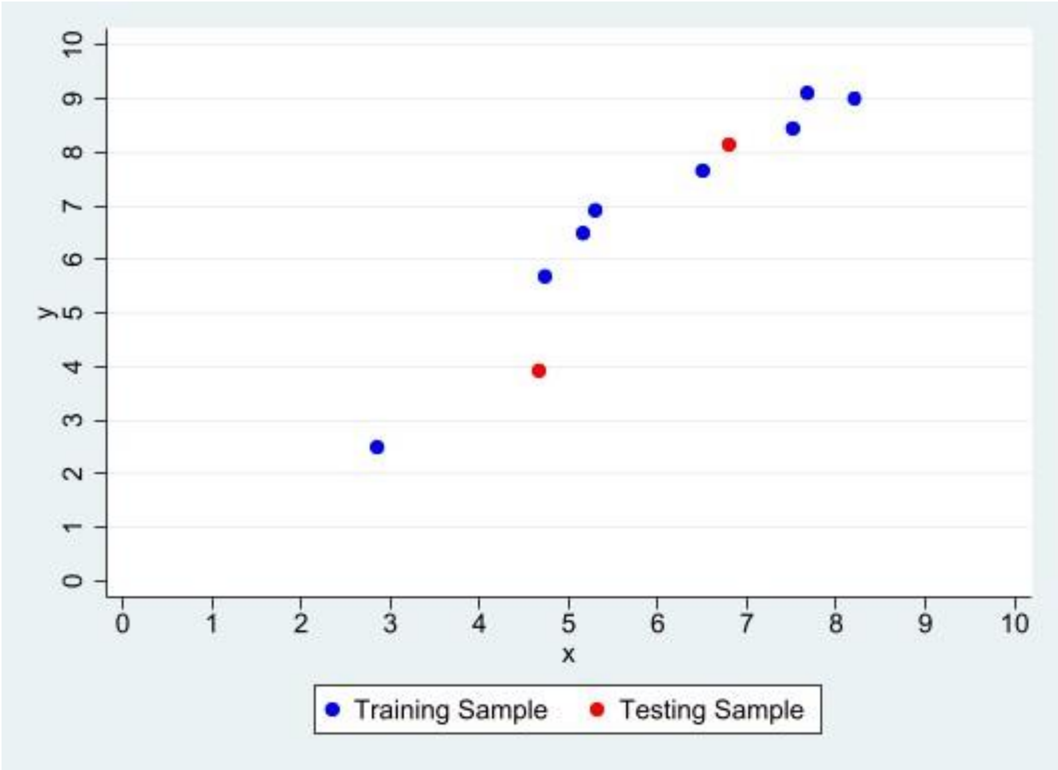
Choosing regularization parameter

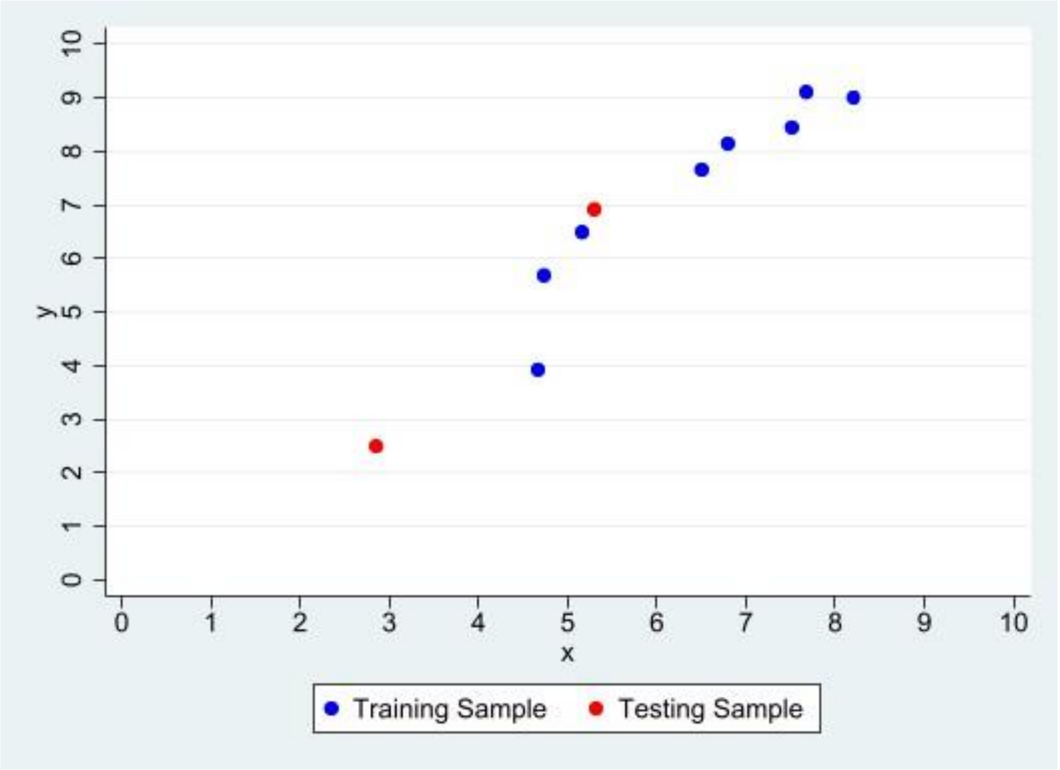
- Typically, data are sparse → Cross-validation

1. Split training data, e.g., 5 sets with 80% training + 20% testing

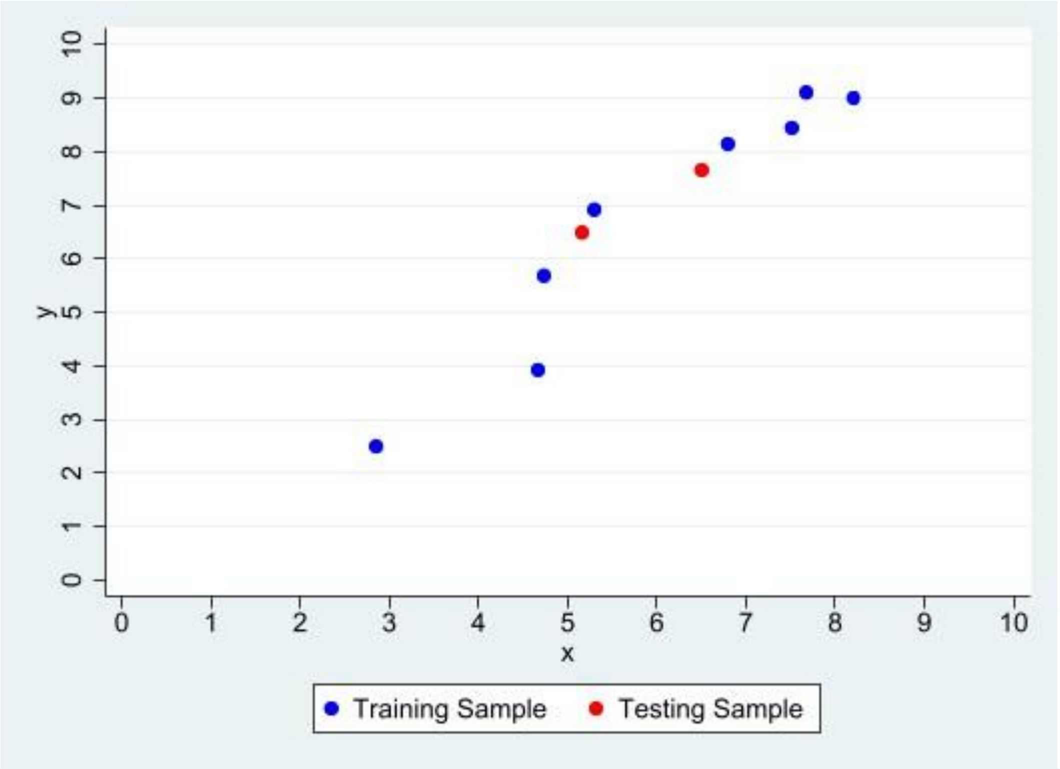




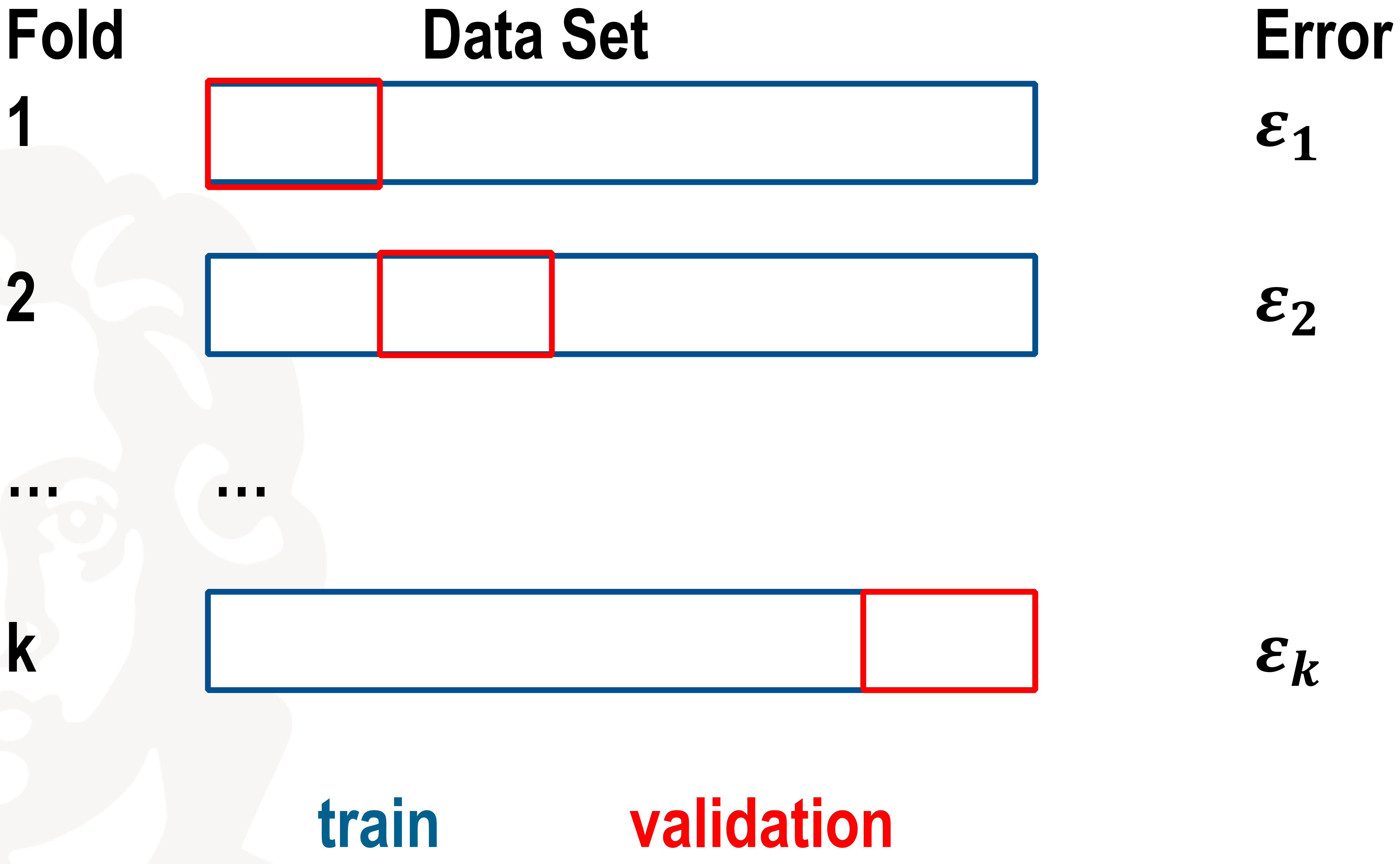




2. Determine optimal parameter(s)
   - In our example: simple (linear) or complex (quadratic)
   - Lasso and Ridge: λ



k-fold cross-validation



Fold

Data Set

Error

1 ε_1

2 ε_2

… …

k ε_k

train validation

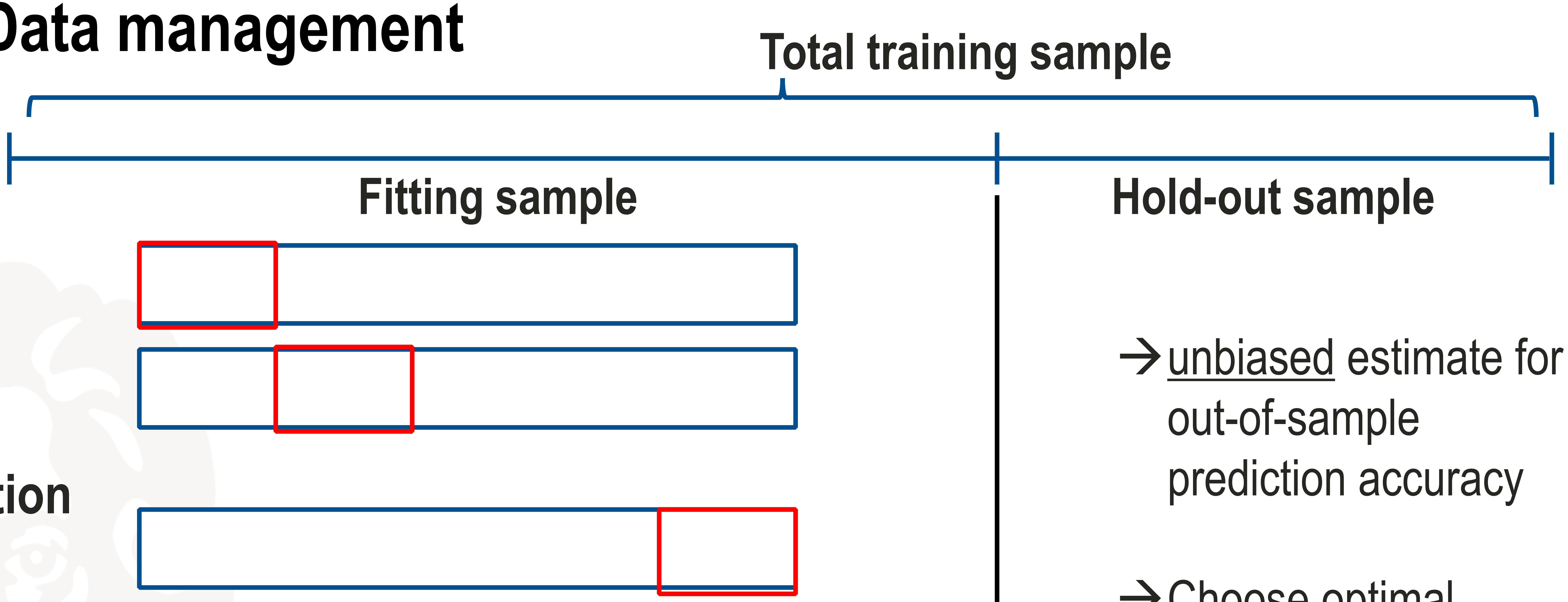
Approach

1. For each λ, compute cross-validation error $$= \frac{\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_k}{k}$$
2. Pick the λ with the lowest error!

Does the cross-validation error tell you how well your function will do out-of-sample?
→Probably too optimistic!

Machine Learning (13)

Data management



Total training sample

Fitting sample

Hold-out sample

k-fold cross-validation

Tuning:
Selecting optimal level of regularization (λ)

→unbiased estimate for out-of-sample prediction accuracy

→Choose optimal method

Firewall principle

Summary

- Steps in machine learning
  1. Choose flexible functional forms
  2. Regularize: limit their expressiveness
  3. Tune: learn how much to regularize from the data

- Model combinations
  - $\hat{f}(x) = w_1 \hat{f}_1 + w_2 \hat{f}_2 + \cdots + w_K \widehat{f}_K$
  - How to choose weights? → cross-validation
  - Combinations typically outperform the best single predictor.

Coming back to Antweiler and Frank (2004)

- A & F (2004) not really a machine learning approach
  - No regularization, no tuning (because of Naïve Bayes)
  - No hold-out sample
  - Not even cross-validation
- Insufficient information on accuracy.
- Not clear which words drive results.

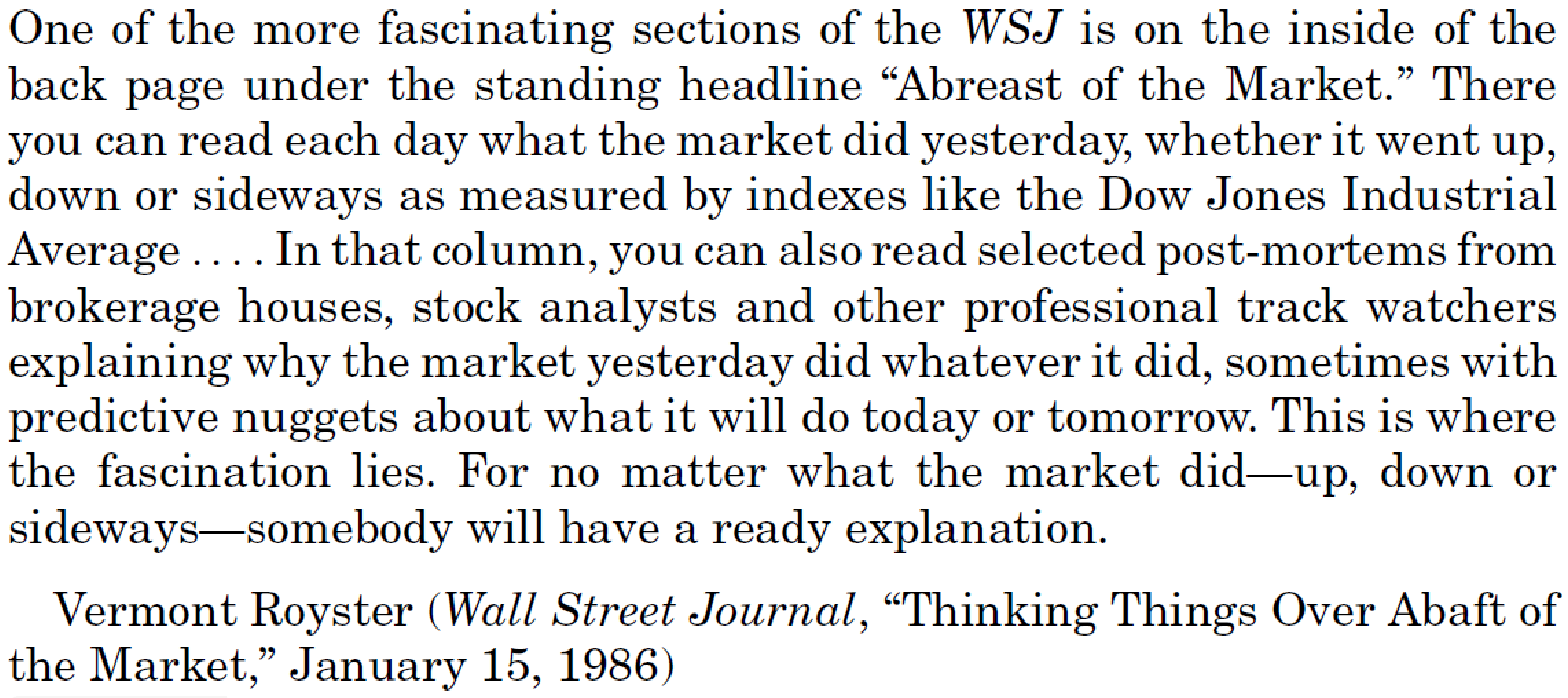
  For LASSO or ridge, one could easily show words with largest (abs.) coefficients.

2. Tetlock (2007)

Tetlock (2007) – Motivation (1)

Motivation of Tetlock (2007)

- ‘Abreast of the Market’ column in the WSJ



> One of the more fascinating sections of the *WSJ* is on the inside of the back page under the standing headline “Abreast of the Market.” There you can read each day what the market did yesterday, whether it went up, down or sideways as measured by indexes like the Dow Jones Industrial Average . . . . In that column, you can also read selected post-mortems from brokerage houses, stock analysts and other professional track watchers explaining why the market yesterday did whatever it did, sometimes with predictive nuggets about what it will do today or tomorrow. This is where the fascination lies. For no matter what the market did—up, down or sideways—somebody will have a ready explanation.
>
> Vermont Royster (*Wall Street Journal*, “Thinking Things Over Abaft of the Market,” January 15, 1986)

- What is the relation between the content of the ‘Abreast of the Market’ column and daily stock market activity?

Tetlock (2007) – Motivation (2)

Abreast of the Market; The Wall Street Journal; January 7, 2004; 675 words

Title: Sun Microsystems, Brocade Rise; Gateway Loses Large-Cap Status

By Karen Talley, Dow Jones Newswires

- NEW YORK -- Sun Microsystems and Brocade Communications Systems helped the Nasdaq Composite Index hit a two-year high, while the Dow Jones Industrial Average pulled back a bit.
- The Nasdaq gained 10.01 points, or 0.49%, to 2057.37, its highest level in 24 months. The Dow Jones Industrial Average fell 5.41 points, or 0.05%, to 10538.66 after a 134-point rise on Monday, and the S&P 500 index rose 1.45 points, or 0.13%, to 1123.67, a new 20-month high.
- The generally upbeat movement came despite some downbeat economic news. But investors are looking farther out "and buying on what they believe will be an improving economic picture," said Mark Donahoe, managing director, institutional sales trading, at Piper Jaffray. "We're starting to see much more institutional involvement."
- Sun Microsystems gained 33 cents, or 7%, to $5.03 after Merrill Lynch raised its sales and earnings estimates, saying checks, though not complete, suggest the maker of large computer systems experienced a strong close to the latest quarter.

Tetlock (2007) – Motivation (3)

Abreast of the Market; The Wall Street Journal; January 7, 2004 – continued

- Brocade rose 33 cents, or 5.2%, to 6.63 after Lehman Brothers upgraded the computer-storage-switch supplier to overweight from equal-weight on expectations of higher revenue and margin expansion.
- Despite the Dow's fall, there were some standouts. J.P. Morgan Chase was among the blue-chip average's best percentage performers, rising 92 cents, or 2.5%, to 37.47 after Prudential Equity Group raised shares to overweight from neutral, saying it expects 2004 earnings to beat expectations. […]
- There were some severe laggards, though, including Gateway, which plunged 64 cents, or 13%, to 4.34 after saying it expects a fourth-quarter loss of nine cents to 15 cents a share before items, when analysts were looking for a loss of 12 cents. With the decline, Gateway became a small stock because its market capitalization fell below the $1.5 billion minimum this column uses as a cutoff for large-cap stocks.

→ Article does not contain new information. It is a description of what happened in the stock market yesterday.

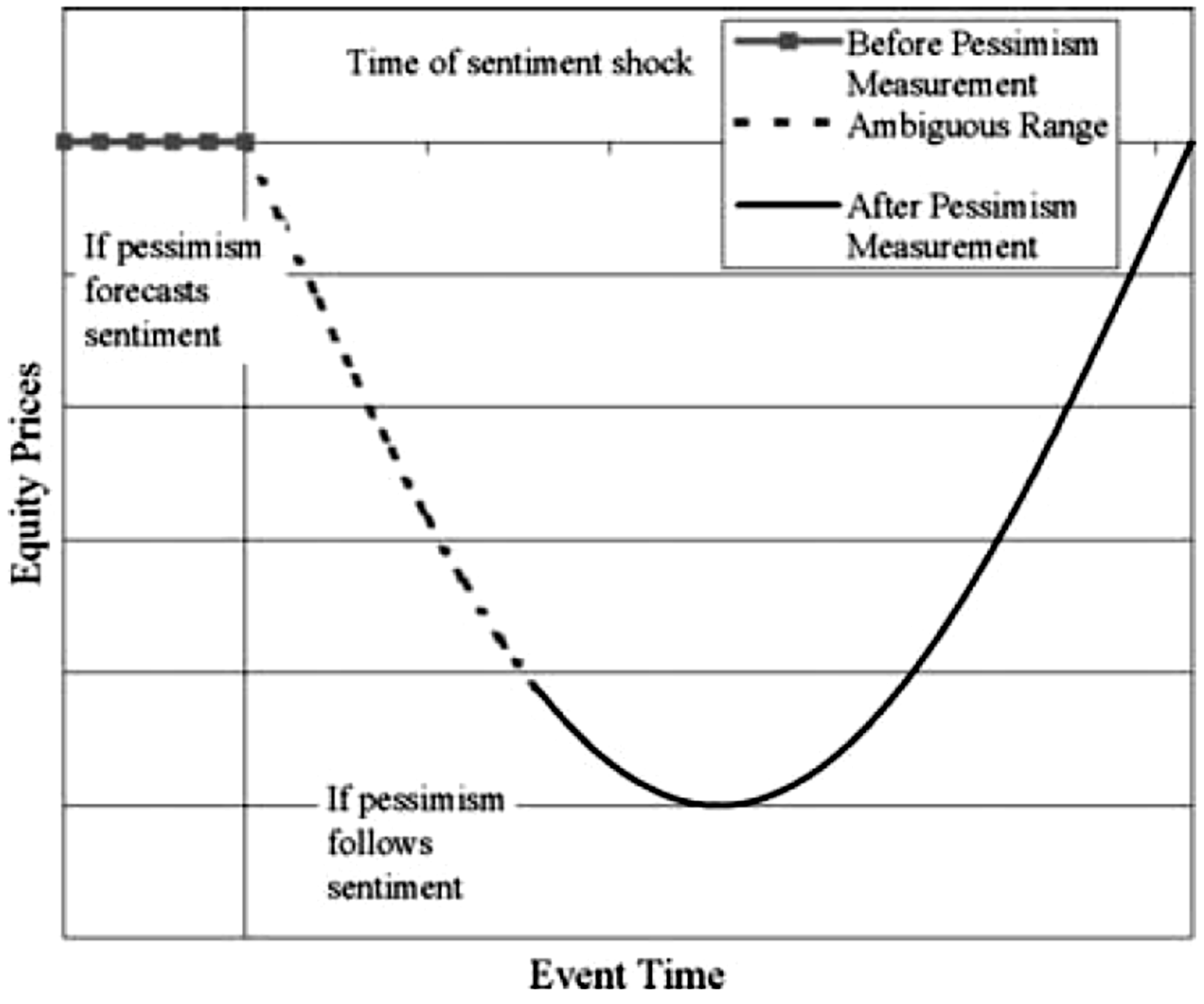
→ In efficient markets, the verbal information should not allow for return prediction.

→ However, if (some) investors are sentiment traders, the tone of the column may have an impact.

Tetlock (2007) – Theoretical background

Theoretical background and economic story

Relation between media pessimism and sentiment; Tetlock (2007) – Figure 1



- Tetlock considers “Abreast of the Market”-tone to be investor sentiment.
- “Investor sentiment” = belief that is not justified by fundamentals.
- Any deviation from fundamental value will be corrected by rational arbitrageurs at some point.

Potential empirical outcomes:

1. Media pessimism forecasts sentiment
   → return impact and reversal of similar magnitude.
2. Media pessimism follows sentiment
   → return impact smaller than reversal.

Tetlock (2007) - Sample

Data

- ‘Abreast of the Market’ Column
  Probably from Factiva or directly from Dow Jones Newswire.
- Sample period: January 1, 1984 to September 17, 1999.
  → 3,709 trading days
- Capital market data
  - Return of the Dow Jones Industrial Average index.
  - (detrended) NYSE trading volume.

Tetlock (2007) – Tone Measurement (1)

Tone measurement

‘Bag of the word approach’ / dictionary approach:

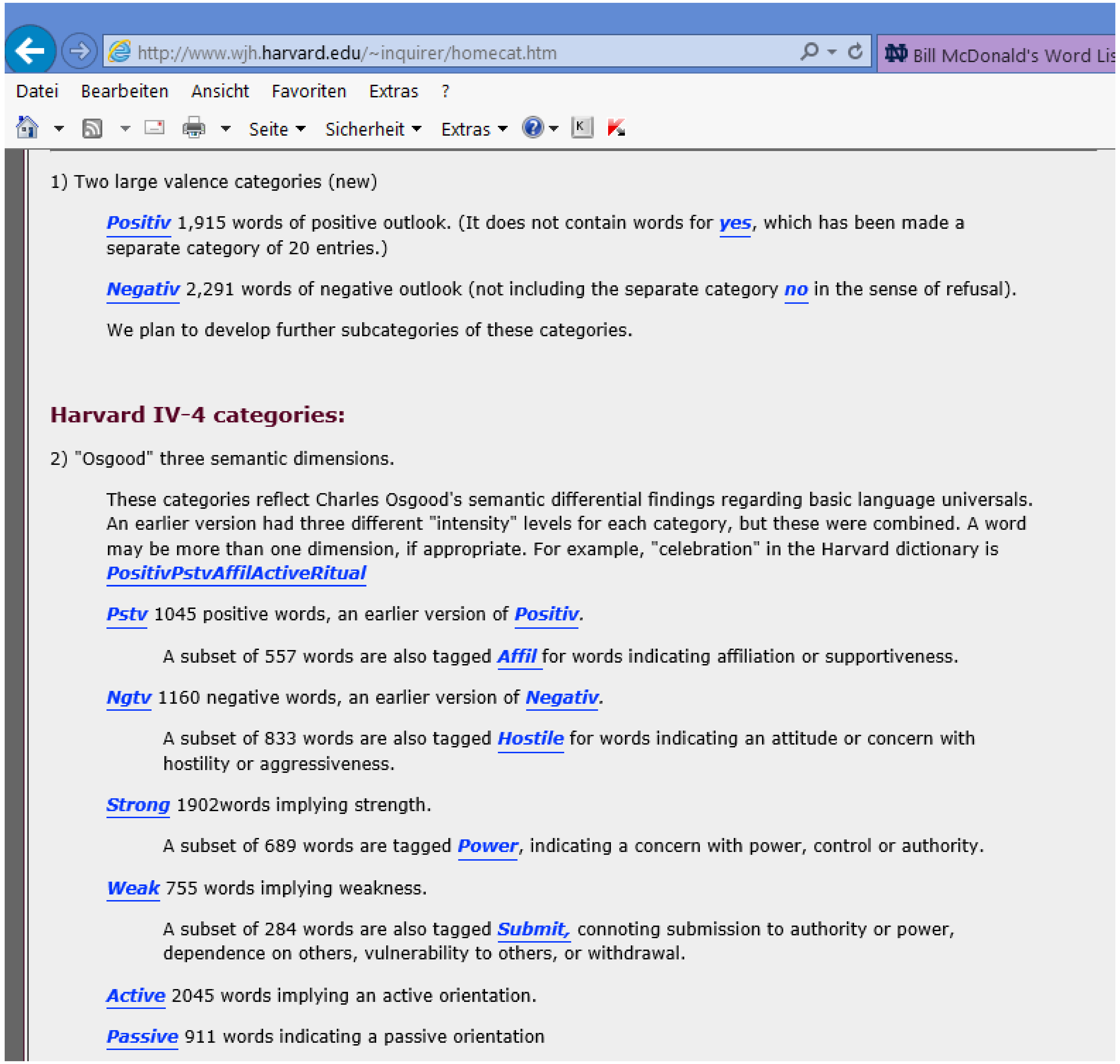
- Count the number of words of a specific category/list (e.g., negative, positive).
- Calculate the fraction of these words by dividing the category word count by the total number of words.

Which word lists?

→ General Inquirer Harvard IV-4 psychosocial dictionary

- 77 dictionaries, e.g.
  - Negative: 2,291 words
  - Positive: 1,902 words
  - Passive: 911 words
  - Pleasure: 168 words
- The dictionaries are available at: http://www.wjh.harvard.edu/~inquirer/homecat.htm

Tetlock (2007) – General Inquirer



1) Two large valence categories (new)

Positiv 1,915 words of positive outlook. (It does not contain words for **yes**, which has been made a separate category of 20 entries.)

Negativ 2,291 words of negative outlook (not including the separate category **no** in the sense of refusal).

We plan to develop further subcategories of these categories.

Harvard IV-4 categories:

2) "Osgood" three semantic dimensions.

These categories reflect Charles Osgood's semantic differential findings regarding basic language universals. An earlier version had three different "intensity" levels for each category, but these were combined. A word may be more than one dimension, if appropriate. For example, "celebration" in the Harvard dictionary is **PositivPstvAffilActiveRitual**

Pstv 1045 positive words, an earlier version of **Positiv**.

A subset of 557 words are also tagged **Affil** for words indicating affiliation or supportiveness.

Ngtv 1160 negative words, an earlier version of **Negativ**.

A subset of 833 words are also tagged **Hostile** for words indicating an attitude or concern with hostility or aggressiveness.

Strong 1902words implying strength.

A subset of 689 words are tagged **Power**, indicating a concern with power, control or authority.

Weak 755 words implying weakness.

A subset of 284 words are also tagged **Submit,** connoting submission to authority or power, dependence on others, vulnerability to others, or withdrawal.

Active 2045 words implying an active orientation.

Passive 911 words indicating a passive orientation

Tetlock (2007) – Tone Measurement (2)

Tone measurement

How to aggregate the 77 dimensions into a single factor?

→ Principal component analysis (PCA).

- ○ Linear combination of the General Inquirer categories.
- ○ Choose the factor with the greatest variance.

- Results of the PCA:
  - ○ Positive weight: negative, weak, fail, and fall categories.
  - ○ Negative weight: positive category.

  → first factor is a pessimism factor.

- Tetlock (2007) uses the pessimism factor as well as the negative and weak categories.

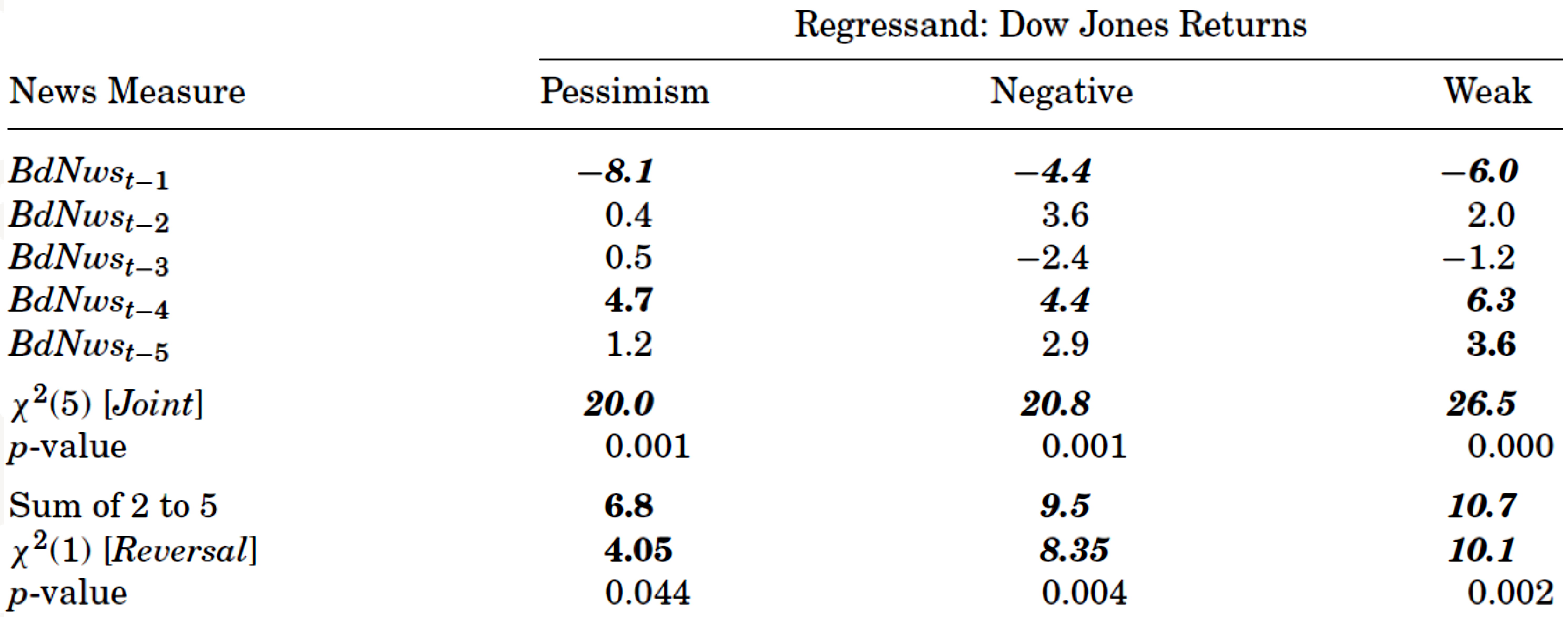
Tetlock (2007) – Sentiment and DJIA Returns (1)

Main result – Sentiment and market returns

Time-series regressions of returns on sentiment; Tetlock (2007) - Table 2

$$Dow_t = \alpha_1 + \beta_1 \cdot L5(Dow_t) + \gamma_1 \cdot L5(BdNws_t) + \delta_1 \cdot L5(Vlm_t) + \lambda_1 \cdot Exog_{t-1} + \varepsilon_{1t}$$

- Exog.: January dummy, day-of-the-week dummies, October 19, 1987 dummy.
- Coefficients measure the effect of a one std. dev. increase in negative investor sentiment on returns (in bp).



	Regressand: Dow Jones Returns		
News Measure	Pessimism	Negative	Weak
$BdNws_{t-1}$	***−8.1***	***−4.4***	***−6.0***
$BdNws_{t-2}$	0.4	3.6	2.0
$BdNws_{t-3}$	0.5	−2.4	−1.2
$BdNws_{t-4}$	**4.7**	***4.4***	***6.3***
$BdNws_{t-5}$	1.2	2.9	**3.6**
$\chi^2(5)$ [*Joint*]	***20.0***	***20.8***	***26.5***
p-value	0.001	0.001	0.000
Sum of 2 to 5	**6.8**	***9.5***	***10.7***
$\chi^2(1)$ [*Reversal*]	**4.05**	***8.35***	***10.1***
p-value	0.044	0.004	0.002

- Low sentiment predicts low market returns the next day.
- Return reversal on the subsequent four days is about the same magnitude as initial reaction. → media tone predicts sentiment.

Tetlock (2007) - Conclusion

Conclusion

- Tone of the popular ‘Abreast of the Market’ column predicts stock market returns.
- Finding are consistent with noise trader model of DeLong et al. (1990a) and liquidity trader model of Campbell et al. (1993).
  → content of the WSJ column does not contain fundamental information but predicts investor sentiment.
- Trading volume increases after high and low sentiment.

Tone measurement

- Simply counting words
  - seems to work well
  - and is easily understood (no black box).
- Negative and weak words seem to be a good choice to measure tone.
- Is the Harvard dictionary a good choice in a business context? → next paper.

3. Loughran and McDonald (2011)

Loughran and McDonald (2011) - Motivation

Is the Harvard dictionary suitable for a business context?

Analyzing the words in the dictionary shows

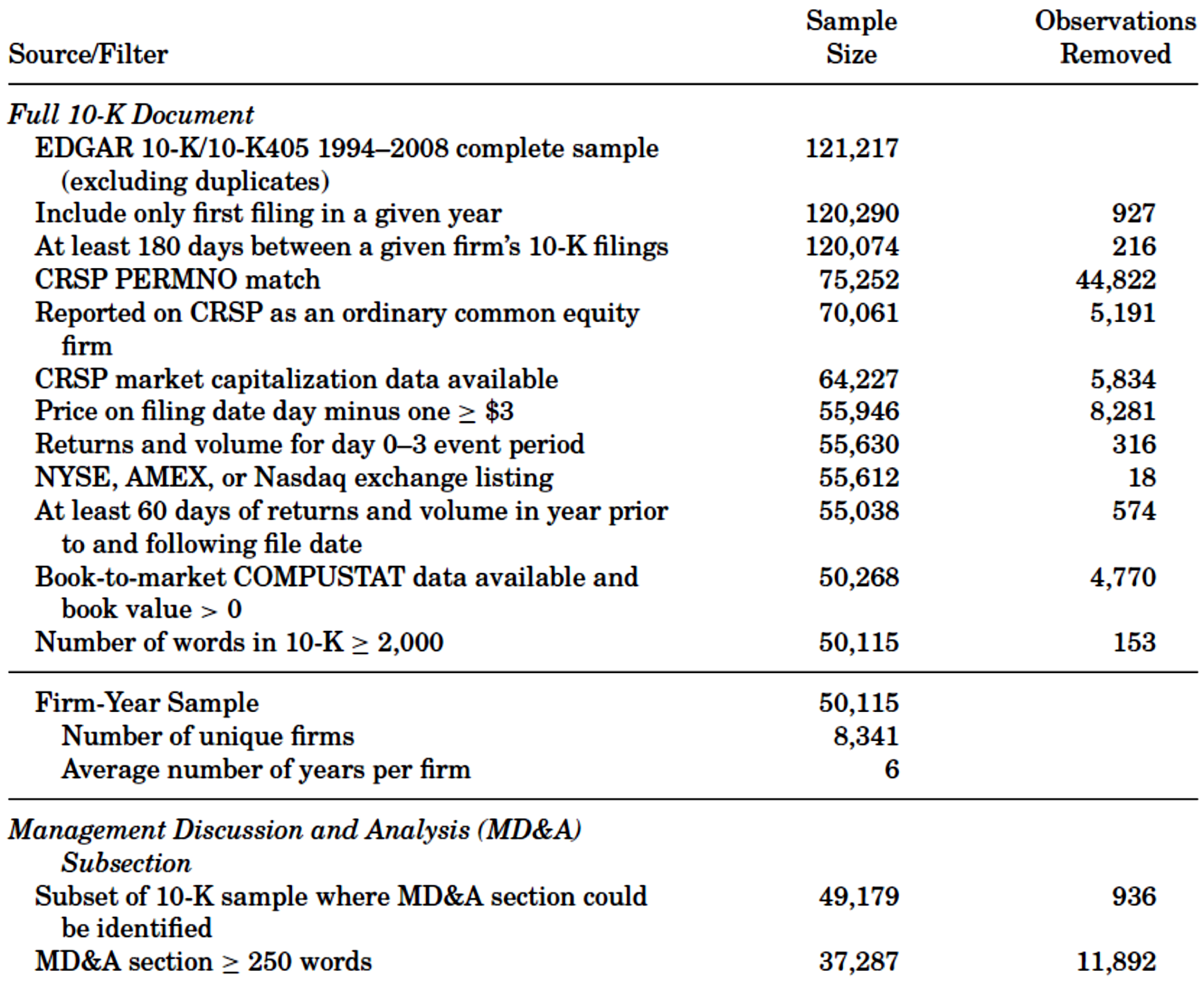
- Neutral meaning
  - Examples: tax, costs, expense, liabilities.
  - → tone measurement is noisy.
- Systematic bias
  - Capital → banking and insurance
  - Crude → oil industry
  - Mine → precious metals and coal
  - Illustration of the magnitude of the problem: in the 1999 10-K of Coeur d’Alene Mines Corporation, the word ‘mine’ accounts for 25% of all negative words.

Main result of the study: almost 75% of the words in the Harvard IV psychosocial dictionary are misclassified in business contexts.

Loughran and McDonald (2011) – Sample

Sample

- Sample period 1994 to 2008.
- All form 10-K filings available at the Securities and Exchange Commission (SEC).
- Excluding firms with missing
  - ○ stock market information
    → not in CRSP (traded over the counter, private companies, REITs).
  - ○ accounting information
    → not in Compustat.



Source/Filter	Sample Size	Observations Removed
Full 10-K Document		
EDGAR 10-K/10-K405 1994–2008 complete sample (excluding duplicates)	121,217	
Include only first filing in a given year	120,290	927
At least 180 days between a given firm's 10-K filings	120,074	216
CRSP PERMNO match	75,252	44,822
Reported on CRSP as an ordinary common equity firm	70,061	5,191
CRSP market capitalization data available	64,227	5,834
Price on filing date day minus one ≥ $3	55,946	8,281
Returns and volume for day 0–3 event period	55,630	316
NYSE, AMEX, or Nasdaq exchange listing	55,612	18
At least 60 days of returns and volume in year prior to and following file date	55,038	574
Book-to-market COMPUSTAT data available and book value > 0	50,268	4,770
Number of words in 10-K ≥ 2,000	50,115	153
Firm-Year Sample	50,115	
Number of unique firms	8,341	
Average number of years per firm	6	
Management Discussion and Analysis (MD&A) Subsection		
Subset of 10-K sample where MD&A section could be identified	49,179	936
MD&A section ≥ 250 words	37,287	11,892

Loughran and McDonald (2011) – Word lists (1)

Loughran and McDonald’s word lists

1. Negative: 2,337 words
   - 1,121 overlap with Harvard negative
   - Restated, litigation, termination, unpaid, investigation, serious, deterioration, etc.
2. Positive: 353 words
   - Achieve, efficient, improve, profitable, etc.
3. Uncertainty: 285 words
   - General notion on imprecision, not only risk
   - Approximate, depend, fluctuate, indefinite, uncertain, etc.
4. Litigious: 731 words
   - Claimant, deposition, testimony, etc.
5. Modal strong: 19 words
   - Always, highest, must, etc.
6. Modal weak: 27 words
   - Could, depending, might, etc.

Loughran and McDonald (2011) – Word lists (2)

Details on the construction of the dictionaries

- How are these lists created?
  1. Take the list of all words contained in the 10-Ks.
  2. Manually classify all words that occur in at least 5% of the filings.

- Word lists can be downloaded from Bill McDonald’s webpage http://www3.nd.edu/~mcdonald/Word_Lists.html

- List include inflected versions of the word lists
  - Accident, accidental, accidentally, and accidents
  - The expand the original Harvard negative list from 2,005 (word stem) to 4,187 words (incl. inflections)
  - Problem with stemming: odd vs. odds, good vs. goods (costs of goods sold).

- Loughran and McDonald control for negation when using the positive word list
  - Simple negation is taken into account. → Check whether one of the six words (*no, not, none, neither, never, nobody*) occurs within three words preceding a positive word.

Loughran and McDonald (2011) – Tone measurement (1)

Details on tone measurement

- Tone is measured by a simple word count
  1. Count the number of XXX words.
  2. Count the number of total words.
  3. Calculate the fraction of XXX words.

- Alternatively, a weighting scheme is applied.
  - tf.idf = term frequency and inverse-document frequency

$$w_{i,j} = \begin{cases} \dfrac{(1+\log(tf_{i,j}))}{(1+\log(a_j))} \log \dfrac{N}{df_i} & \text{if } \ tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

  - $w_{i,j}$ weight of word i in document j
  - N total number of 10-Ks
  - df_i number of 10-Ks in which word i is found
  - $tf_{i,j}$ raw number of word i in document j
  - a_j average word count in the document j

Example of tf.idf weighting

- "recession" is found 2 times in the 10-K ($tf_{i,j} = 2$).
- The 10-K contains 400 words in total and 200 unique words ($a_j = \frac{400}{200} = 2$).
- There are 20 10-Ks in total ($N = 20$).
- In 4 of them, "recession" is found ($df_i = 4$).

→ $w_{recession,10-K} = \frac{(1+\log(2))}{(1+\log(2))} \log \frac{20}{4}$

$= 1.6094$

- "loss" is found 10 times in the 10-K ($tf_{i,j} = 10$).
- In 18 10-Ks, "loss" is found ($df_i = 18$).

→ $w_{sales,10-K} = \frac{(1+\log(10))}{(1+\log(2))} \log \frac{20}{18}$

$= 0.2055$

There are different ways for computing term frequency (e.g., without the denominator $(1+\log(a_j))$ or using the standard word count).

Loughran and McDonald (2011) – Tone measurement (2)

Details on tone measurement - continued

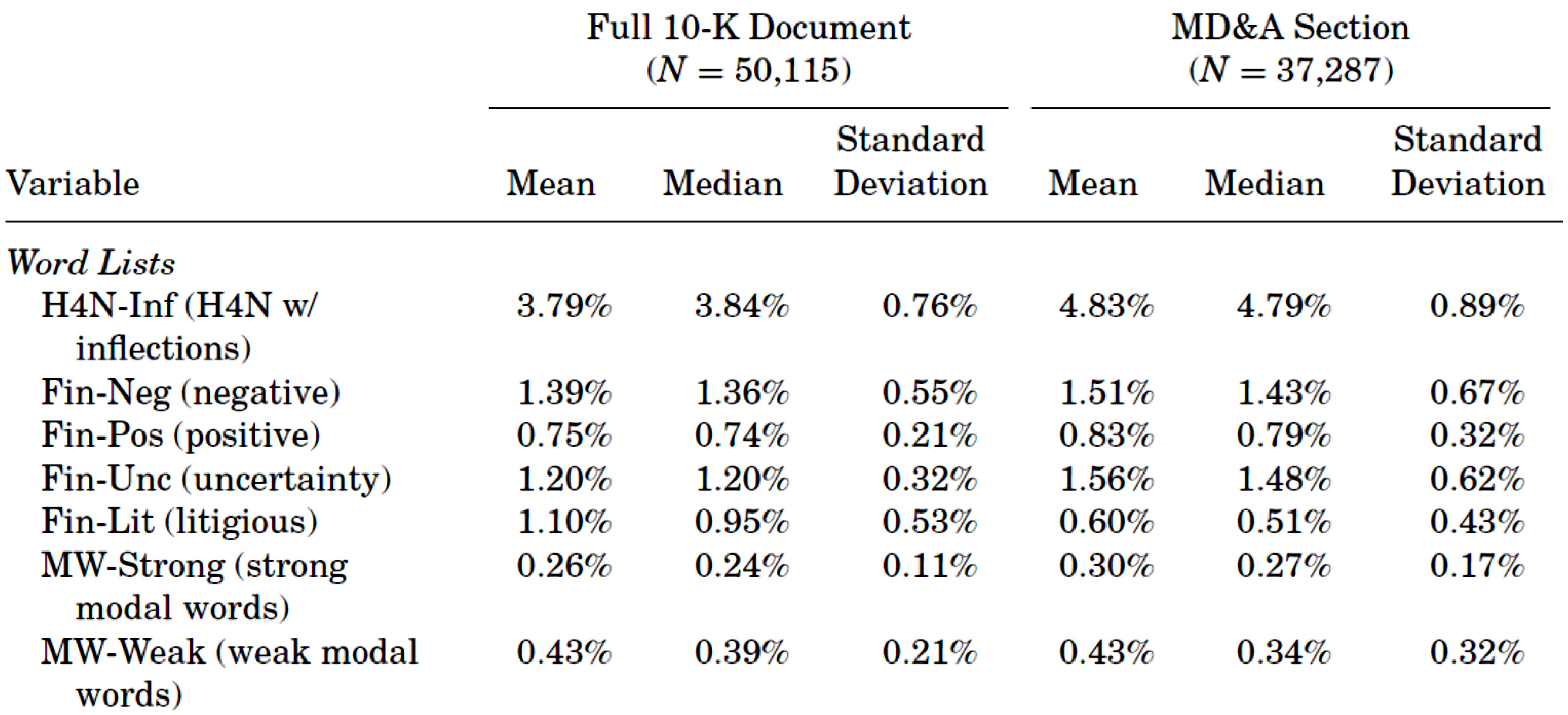
Weighting scheme

- tf.idf $$w_{i,j} = \begin{cases} \frac{(1+\log(tf_{i,j}))}{(1+\log(a_j))} \log \frac{N}{df_i} & \text{if } \; tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

- The first term attenuates the impact of high frequency words with a log transformation.
  Loss appears 1.79 million times; aggravates appears 10 times.
- The second term modifies the impact based on its commonality.
  Loss appears in more than 90% of the documents.

Loughran and McDonald (2011) – Summary statistics

Summary statistics



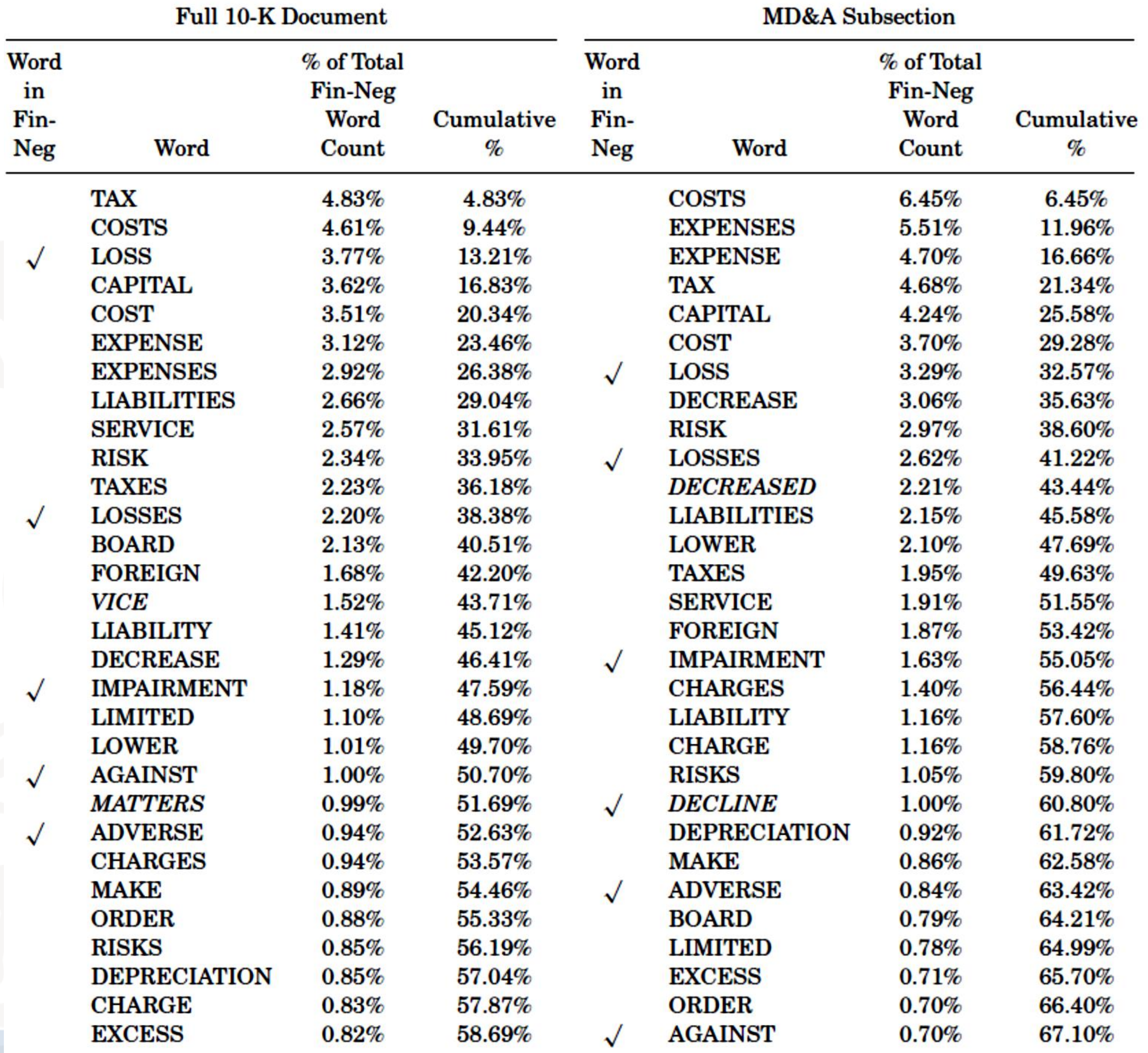
	Full 10-K Document (*N* = 50,115)			MD&A Section (*N* = 37,287)		
Variable	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation
Word Lists						
H4N-Inf (H4N w/ inflections)	3.79%	3.84%	0.76%	4.83%	4.79%	0.89%
Fin-Neg (negative)	1.39%	1.36%	0.55%	1.51%	1.43%	0.67%
Fin-Pos (positive)	0.75%	0.74%	0.21%	0.83%	0.79%	0.32%
Fin-Unc (uncertainty)	1.20%	1.20%	0.32%	1.56%	1.48%	0.62%
Fin-Lit (litigious)	1.10%	0.95%	0.53%	0.60%	0.51%	0.43%
MW-Strong (strong modal words)	0.26%	0.24%	0.11%	0.30%	0.27%	0.17%
MW-Weak (weak modal words)	0.43%	0.39%	0.21%	0.43%	0.34%	0.32%

Loughran and McDonald (2011) –Table 2

- Differences in levels driven by the number of words in the dictionary.
- Higher frequencies in the MD&A may indicate that it is more informative.

Loughran and McDonald (2011) – Comparison of word lists (1)

Most frequent words from the Harvard negative dictionary



Full 10-K Document				MD&A Subsection			
Word in Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %
	TAX	4.83%	4.83%		COSTS	6.45%	6.45%
	COSTS	4.61%	9.44%		EXPENSES	5.51%	11.96%
✓	LOSS	3.77%	13.21%		EXPENSE	4.70%	16.66%
	CAPITAL	3.62%	16.83%		TAX	4.68%	21.34%
	COST	3.51%	20.34%		CAPITAL	4.24%	25.58%
	EXPENSE	3.12%	23.46%		COST	3.70%	29.28%
	EXPENSES	2.92%	26.38%	✓	LOSS	3.29%	32.57%
	LIABILITIES	2.66%	29.04%		DECREASE	3.06%	35.63%
	SERVICE	2.57%	31.61%		RISK	2.97%	38.60%
	RISK	2.34%	33.95%	✓	LOSSES	2.62%	41.22%
	TAXES	2.23%	36.18%		*DECREASED*	2.21%	43.44%
✓	LOSSES	2.20%	38.38%		LIABILITIES	2.15%	45.58%
	BOARD	2.13%	40.51%		LOWER	2.10%	47.69%
	FOREIGN	1.68%	42.20%		TAXES	1.95%	49.63%
	VICE	1.52%	43.71%		SERVICE	1.91%	51.55%
	LIABILITY	1.41%	45.12%		FOREIGN	1.87%	53.42%
	DECREASE	1.29%	46.41%	✓	IMPAIRMENT	1.63%	55.05%
✓	IMPAIRMENT	1.18%	47.59%		CHARGES	1.40%	56.44%
	LIMITED	1.10%	48.69%		LIABILITY	1.16%	57.60%
	LOWER	1.01%	49.70%		CHARGE	1.16%	58.76%
✓	AGAINST	1.00%	50.70%		RISKS	1.05%	59.80%
	MATTERS	0.99%	51.69%	✓	*DECLINE*	1.00%	60.80%
✓	ADVERSE	0.94%	52.63%		DEPRECIATION	0.92%	61.72%
	CHARGES	0.94%	53.57%		MAKE	0.86%	62.58%
	MAKE	0.89%	54.46%	✓	ADVERSE	0.84%	63.42%
	ORDER	0.88%	55.33%		BOARD	0.79%	64.21%
	RISKS	0.85%	56.19%		LIMITED	0.78%	64.99%
	DEPRECIATION	0.85%	57.04%		EXCESS	0.71%	65.70%
	CHARGE	0.83%	57.87%		ORDER	0.70%	66.40%
	EXCESS	0.82%	58.69%	✓	AGAINST	0.70%	67.10%

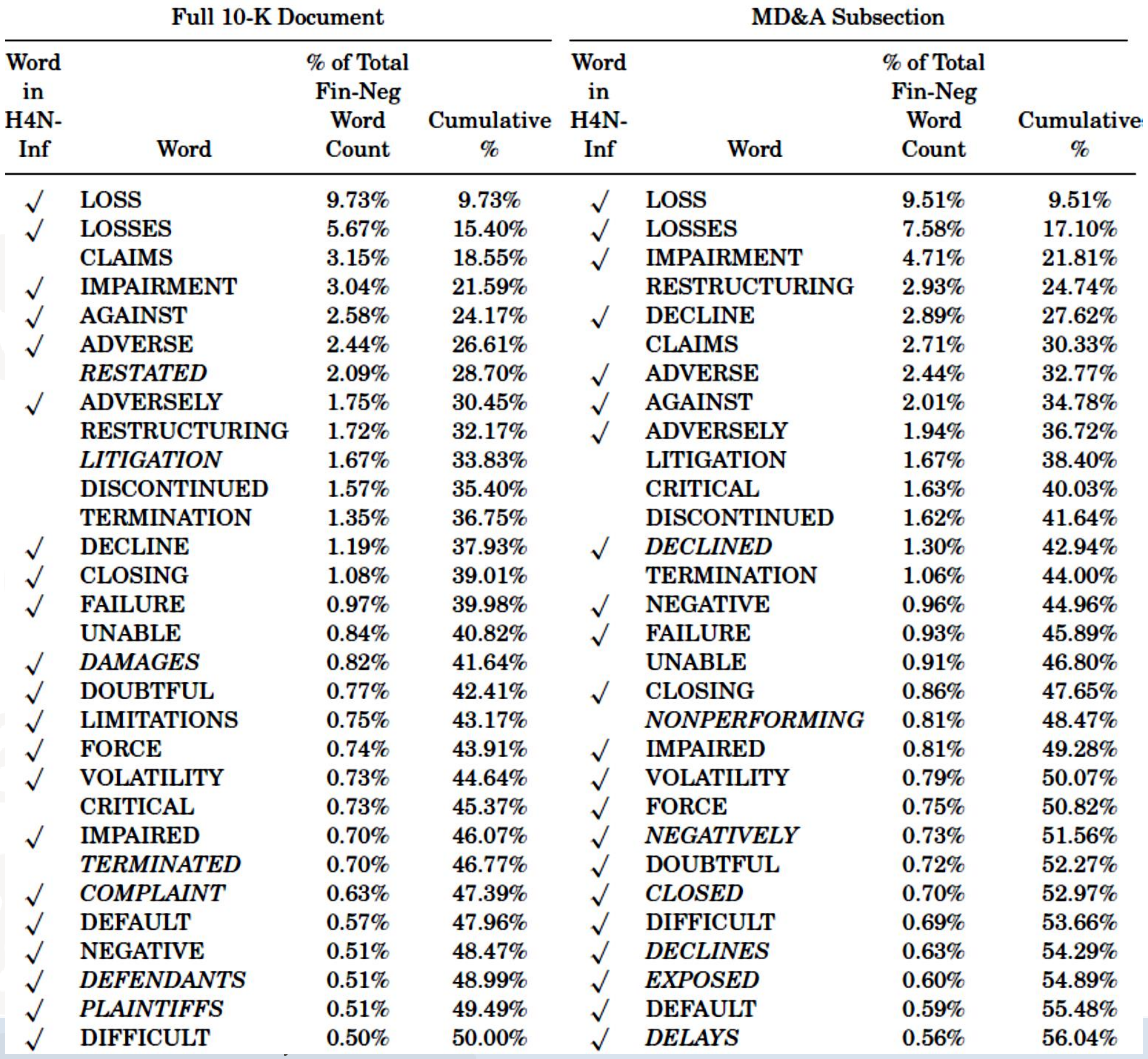
Results

- List is dominated by HVD neg. words that are not meaningful in a business context.
- Only 5 (6) of the 30 most frequent HVD neg. words in the overall text (in the MD&A) are included in LMD neg.

Loughran and McDonald (2011) –Table 3, part 1

Loughran and McDonald (2011) – Comparison of word lists (2)

Most frequent words from the Loughran and McDonald negative dictionary



Full 10-K Document				MD&A Subsection			
Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %
✓	LOSS	9.73%	9.73%	✓	LOSS	9.51%	9.51%
✓	LOSSES	5.67%	15.40%	✓	LOSSES	7.58%	17.10%
	CLAIMS	3.15%	18.55%	✓	IMPAIRMENT	4.71%	21.81%
✓	IMPAIRMENT	3.04%	21.59%		RESTRUCTURING	2.93%	24.74%
✓	AGAINST	2.58%	24.17%	✓	DECLINE	2.89%	27.62%
✓	ADVERSE	2.44%	26.61%		CLAIMS	2.71%	30.33%
	RESTATED	2.09%	28.70%	✓	ADVERSE	2.44%	32.77%
✓	ADVERSELY	1.75%	30.45%	✓	AGAINST	2.01%	34.78%
	RESTRUCTURING	1.72%	32.17%	✓	ADVERSELY	1.94%	36.72%
	LITIGATION	1.67%	33.83%		LITIGATION	1.67%	38.40%
	DISCONTINUED	1.57%	35.40%		CRITICAL	1.63%	40.03%
	TERMINATION	1.35%	36.75%		DISCONTINUED	1.62%	41.64%
✓	DECLINE	1.19%	37.93%	✓	*DECLINED*	1.30%	42.94%
✓	CLOSING	1.08%	39.01%		TERMINATION	1.06%	44.00%
✓	FAILURE	0.97%	39.98%	✓	NEGATIVE	0.96%	44.96%
	UNABLE	0.84%	40.82%	✓	FAILURE	0.93%	45.89%
✓	*DAMAGES*	0.82%	41.64%		UNABLE	0.91%	46.80%
✓	DOUBTFUL	0.77%	42.41%	✓	CLOSING	0.86%	47.65%
✓	LIMITATIONS	0.75%	43.17%		*NONPERFORMING*	0.81%	48.47%
✓	FORCE	0.74%	43.91%	✓	IMPAIRED	0.81%	49.28%
✓	VOLATILITY	0.73%	44.64%	✓	VOLATILITY	0.79%	50.07%
	CRITICAL	0.73%	45.37%	✓	FORCE	0.75%	50.82%
✓	IMPAIRED	0.70%	46.07%	✓	*NEGATIVELY*	0.73%	51.56%
	TERMINATED	0.70%	46.77%	✓	DOUBTFUL	0.72%	52.27%
✓	*COMPLAINT*	0.63%	47.39%	✓	*CLOSED*	0.70%	52.97%
✓	DEFAULT	0.57%	47.96%	✓	DIFFICULT	0.69%	53.66%
✓	NEGATIVE	0.51%	48.47%	✓	*DECLINES*	0.63%	54.29%
✓	*DEFENDANTS*	0.51%	48.99%	✓	*EXPOSED*	0.60%	54.89%
✓	*PLAINTIFFS*	0.51%	49.49%	✓	DEFAULT	0.59%	55.48%
✓	DIFFICULT	0.50%	50.00%	✓	*DELAYS*	0.56%	56.04%

Results

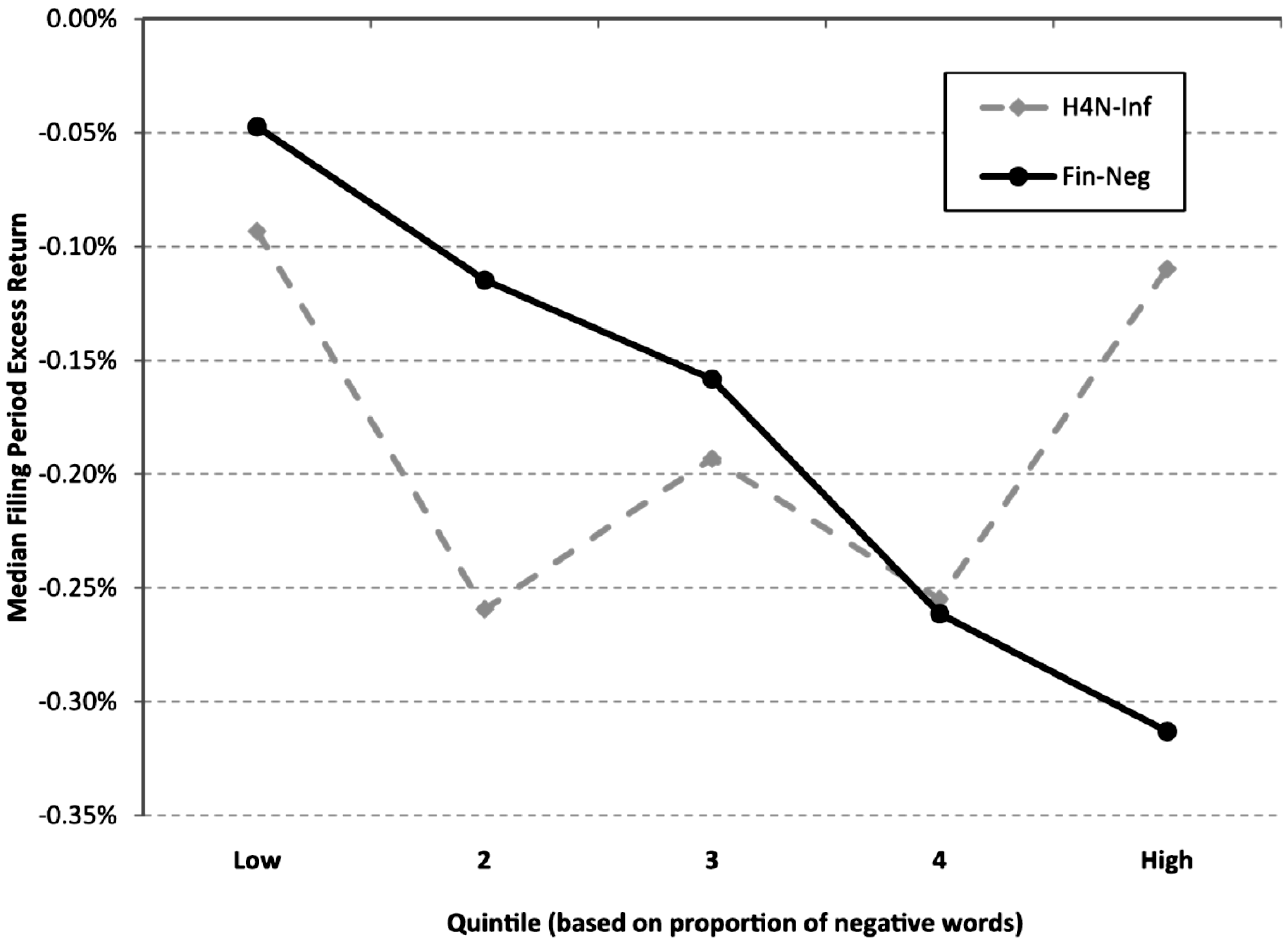
- Words make intuitively sense.
- Large overlap with HVD: only 9 (8) of the 30 most frequent LMD neg. words (in the MD&A) are “new”.

  → LMD neg. is mainly constructed by dropping inappropriate HVD neg. words.

Loughran and McDonald (2011) –Table 3, part 2

Loughran and McDonald (2011) – Comparison of word lists (3)

Relation between HVD/LMD and stock returns



Discussion

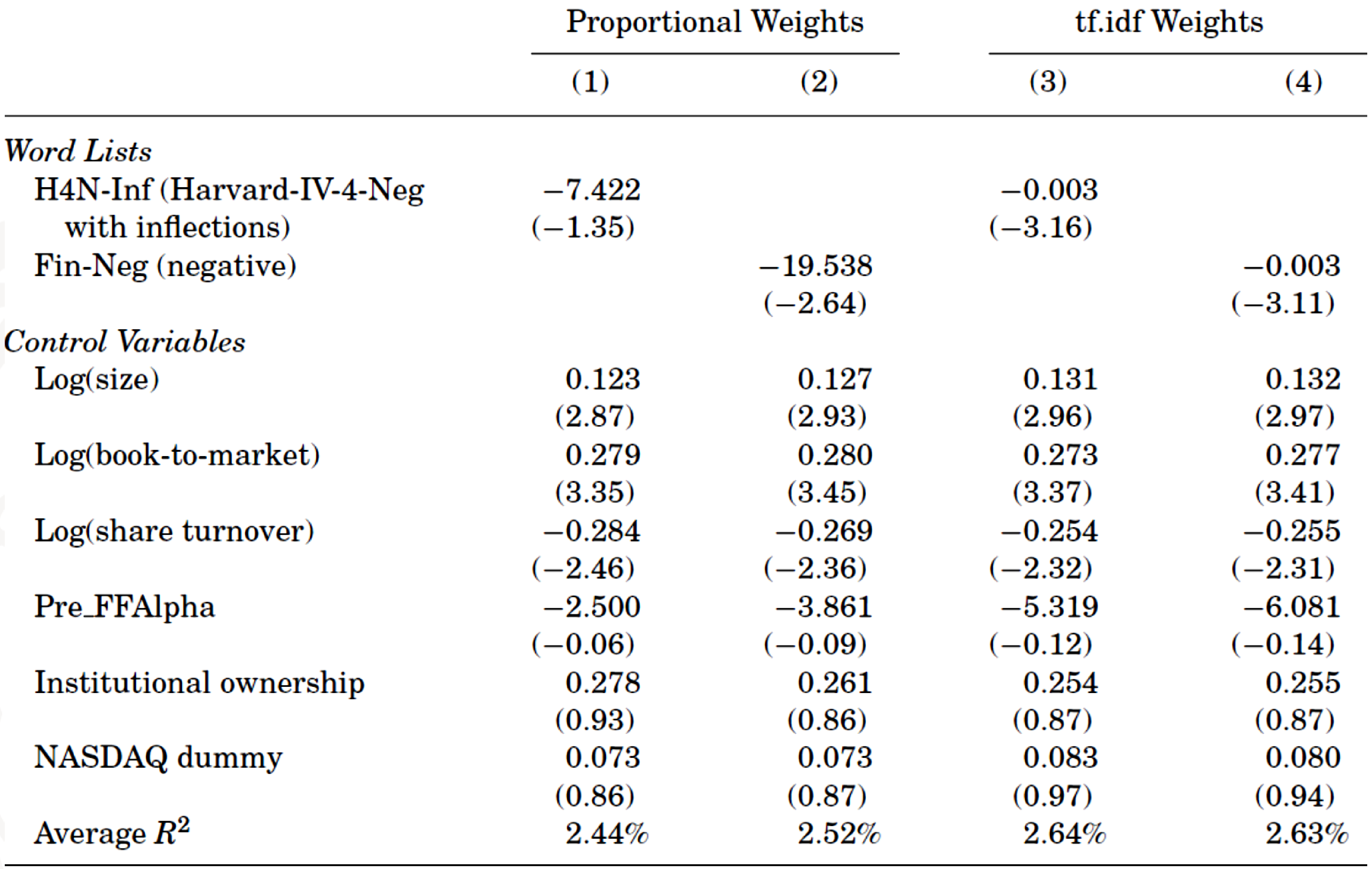
- The figure shows the median 3-day market-excess return around the filing date of tone quintiles.
- As 10-Ks are informative, negativity should be negatively related to returns.
- Result

  While HVD neg. does not show a link to returns, LMD neg. is monotonically related to returns.

Loughran and McDonald (2011) – Figure 1

Loughran and McDonald (2011) – Excess return and document tone (1)

Relation between HVD/LMD and stock returns



	Proportional Weights		tf.idf Weights	
	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf (Harvard-IV-4-Neg with inflections)	−7.422 (−1.35)		−0.003 (−3.16)	
Fin-Neg (negative)		−19.538 (−2.64)		−0.003 (−3.11)
Control Variables				
Log(size)	0.123 (2.87)	0.127 (2.93)	0.131 (2.96)	0.132 (2.97)
Log(book-to-market)	0.279 (3.35)	0.280 (3.45)	0.273 (3.37)	0.277 (3.41)
Log(share turnover)	−0.284 (−2.46)	−0.269 (−2.36)	−0.254 (−2.32)	−0.255 (−2.31)
Pre_FFAlpha	−2.500 (−0.06)	−3.861 (−0.09)	−5.319 (−0.12)	−6.081 (−0.14)
Institutional ownership	0.278 (0.93)	0.261 (0.86)	0.254 (0.87)	0.255 (0.87)
NASDAQ dummy	0.073 (0.86)	0.073 (0.87)	0.083 (0.97)	0.080 (0.94)
Average R^2	2.44%	2.52%	2.64%	2.63%

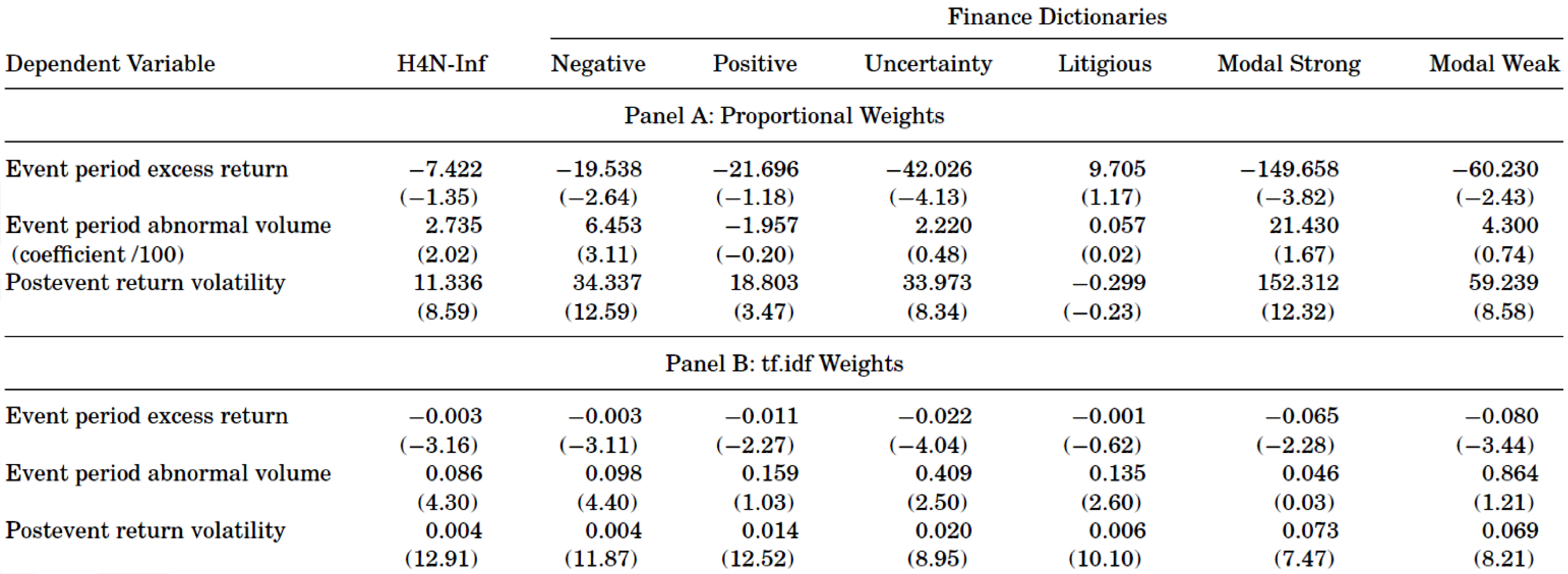
Discussion

- The table shows regressions of 3-day market-excess returns on tone and control variables.
- Results confirm conclusion from previous slide: LMD neg. significantly related to investors' reaction to the 10-K.
- Term-weighted HVD neg. also shows a strong association.
  → tf.idf reduces the weight of the most frequent (and misclassified) words.

Loughran and McDonald (2011) – Table 4

Loughran and McDonald (2011) – Excess return and document tone (2)

Relation between other tone dimensions and stock returns



		Finance Dictionaries					
Dependent Variable	H4N-Inf	Negative	Positive	Uncertainty	Litigious	Modal Strong	Modal Weak
Panel A: Proportional Weights							
Event period excess return	−7.422	−19.538	−21.696	−42.026	9.705	−149.658	−60.230
	(−1.35)	(−2.64)	(−1.18)	(−4.13)	(1.17)	(−3.82)	(−2.43)
Event period abnormal volume	2.735	6.453	−1.957	2.220	0.057	21.430	4.300
(coefficient /100)	(2.02)	(3.11)	(−0.20)	(0.48)	(0.02)	(1.67)	(0.74)
Postevent return volatility	11.336	34.337	18.803	33.973	−0.299	152.312	59.239
	(8.59)	(12.59)	(3.47)	(8.34)	(−0.23)	(12.32)	(8.58)
Panel B: tf.idf Weights							
Event period excess return	−0.003	−0.003	−0.011	−0.022	−0.001	−0.065	−0.080
	(−3.16)	(−3.11)	(−2.27)	(−4.04)	(−0.62)	(−2.28)	(−3.44)
Event period abnormal volume	0.086	0.098	0.159	0.409	0.135	0.046	0.864
	(4.30)	(4.40)	(1.03)	(2.50)	(2.60)	(0.03)	(1.21)
Postevent return volatility	0.004	0.004	0.014	0.020	0.006	0.073	0.069
	(12.91)	(11.87)	(12.52)	(8.95)	(10.10)	(7.47)	(8.21)

Loughran and McDonald (2011) – Table 6

- Table shows regressions of returns (=previous slide), trading volume and volatility on tone.
- LMD run separate regressions for each tone measure → problematic as word lists overlap.
- Some results seem unintuitive. For instance, the negative relation between positivity and returns.

Loughran and McDonald (2011) - Conclusion

Summary and conclusions

- The Harvard dictionaries are not appropriate for analyzing tone in a business context
- Two possible solutions
  - Loughran and McDonald word lists
  - Term weighting

- 'Most important, we show that financial researchers should be cautious when relying on word classification schemes derived outside the domain of business usage. Applying nonbusiness word lists to accounting and finance topics can lead to a high misclassification rate and spurious correlations. <u>All textual analysis ultimately stands or falls by the categorization procedures</u>.' Loughran and McDonald (p.62)

  → **Key take away for your own projects**

Summary of the first section

What have we learnt so far?

- Different approaches: machine learning vs. bag of words.
- Negative words show the strongest association with capital market outcomes.
- The Harvard dictionaries are a good starting point but not appropriate in a business context.
- Qualitative/verbal information matters.

Agenda

- Main contribution on textual analysis in finance
  - Naïve Bayes approach: Antweiler and Frank (2004)
  - Introduction to machine learning
  - Dictionary approach: Tetlock (2007) and Loughran and McDonald (2011)
- Recommendations for your textual analysis
- Selected topics and papers
  - Readability: Loughran and McDonald (2014, 2020)
  - Textual similarity: Tetlock (2011) and Cohen et al. (2020)

Recommendations for your textual analysis

Recommendations (1)

Which dictionary should you use?

- Loughran and McDonald (2011) dictionary has been used in different contexts.
    - Central bank speeches: Schmeling and Wagner (2019).
    - Earnings conference calls: Davis et al. (2015), Dzieliński et al. (2018).
    - IPO prospectuses: Loughran and McDonald (2013).
    - Mutual fund shareholder letters: Hillert et al. (2020).
    - Newspaper articles: Garcia (2013), Hillert et al. (2014).

→ Has become the work horse for textual analysis research in accounting, finance, and economics.

Recommendations (2)

What about other languages?

- Loughran and McDonald (2011) dictionary is available in German.
  - Bannier et al. (2019) develop a German translation of the LMD (2011) dictionary.
  - Using German and English quarterly and annual reports of German companies, they show that the original LMD and their German translation are equivalent.
  - Word list available at: https://www.uni-giessen.de/fbz/fb02/forschung/research-networks/bsfa/textual_analysis/index_html

Recommendations (3)

Checking for plausibility

- List the most frequent words in your categories of interest (positive, negative, etc.).
- Do these words pass a “smell test”?
  - Particularly important for machine learning approaches.
  - Example: Purda and Skillicorn (2015)
    - Goal: predict accounting fraud based on 10-Ks.
    - Most predictive words include, for example, at, as, it, or, on, and may.
    - Economically not very intuitive.
- Depending on the context adjust word list.
- Example: earnings conference calls.
  - LMD positive: “good” → “Good morning”, “That’s a good question”.
  - LMD negative: “question” → “The next questions comes…”, “Thanks for this question.”

Recommendations (4)

Recommendations for editing texts

- Remove single character words → not meaningful anyway.
- Remove numbers.
  Except you are interest in measuring the amount of hard and soft information (e.g., Zhou (2018)).
- Comprehensively account for negations.
  - Approach by LMD (2011) is good starting point.
  - Long forms vs. contracted forms: “are not” vs. “aren’t”, “could not” vs. “couldn’t”.
  - Spelling: “can not” vs. “cannot”.

Recommendations (5)

Recommendations for editing texts - continued

- Stemming
  - Higher precision (inflected version) vs. easier interpretation (stemmed).
  - No dominant approach.
  - Personally, I prefer not to stem the text.
- Stop words
  - Definition: words one does not want to consider in the textual analysis.
    - Very frequent words that are usually uninformative like “and”, “the”, “or”.
    - Words that have ambiguous meaning, e.g., words expressing irony and/or sarcasm.
  - If stop words are equally distributed across texts, they do not affect results.
  - Loughran and McDonald (2016) write “the elimination or special treatment of stop words is typically not necessary.” (p. 1206).
  - Based on my experience results with and without removal of stop words very similar.

Recommendations (6)

Should you use positive words, negative words or net tone?

- Positive words often carry an ambiguous meaning.
- Real-word example: GM’s 2007 annual report
  - Available at: https://www.sec.gov/Archives/edgar/data/40730/000095012408000921/k23797e10vk.htm
  - “In 2007, the global automotive industry continued to show strong sales and revenue growth.” (p. 48).
  - 2007’s net loss(!): $38,732 million (p. 46).

- Negative words are rarely used in an ambiguous way.

- My and Loughran and McDonald’s recommendation: focus on negative words.

Recommendations (7)

Term weighting

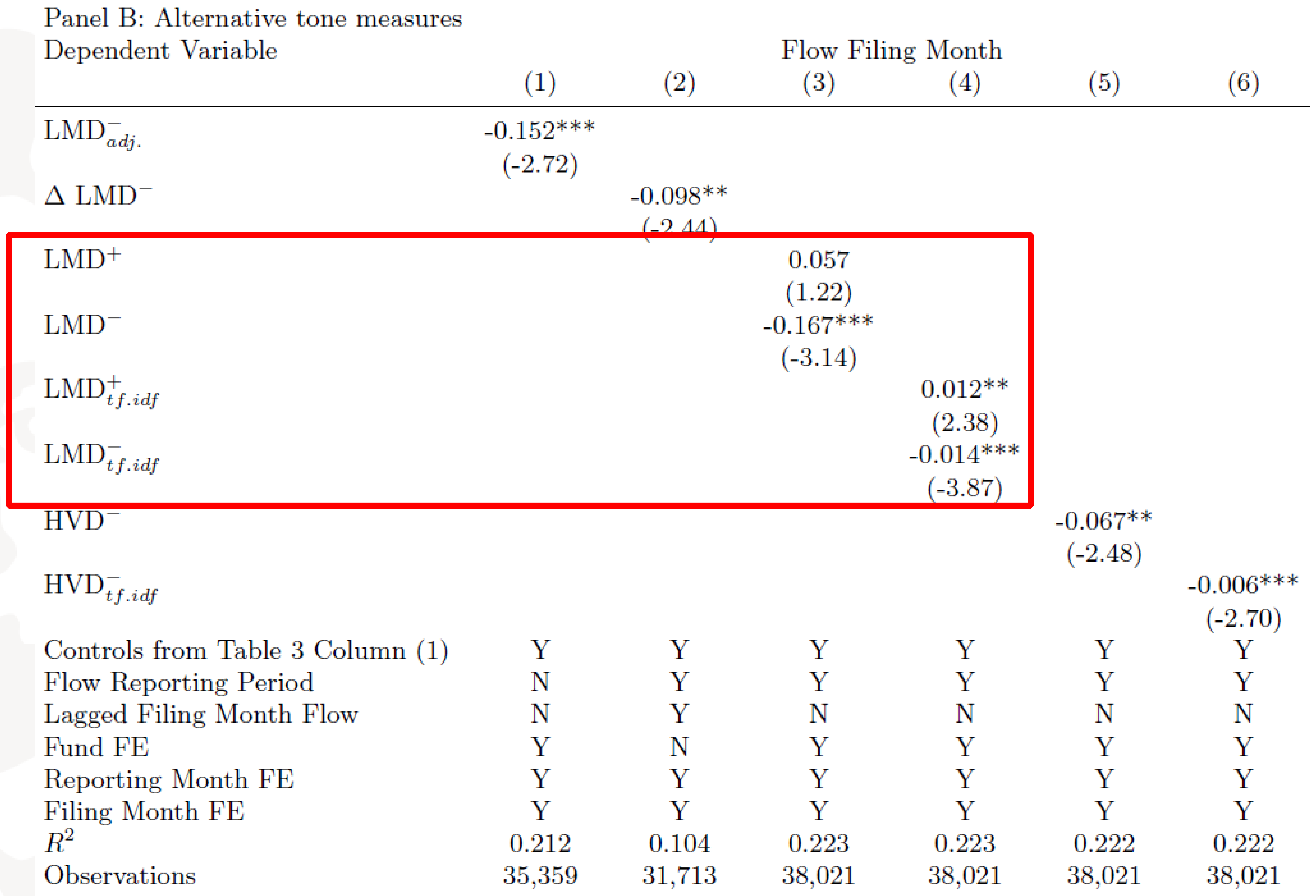
- Term weighting as in Loughran and McDonald (2011; see slides 52 and 53) can be helpful.
- Jegadeesh and Wu (2013) find that term-weighted positive tone is positively related to filing returns of 10-Ks.

  "We also find that the appropriate choice of term weighting in content analysis is at least as important as, and perhaps more important than, complete and accurate compilation of the word list." (p. 712)

- Similar effect in Hillert et al. (2020) (next slide).

Term weighting

- Hillert et al. (2020) find the same for mutual fund shareholder letters (Table 4 – Panel B) Regressions of fund flows on the tone of shareholder letters and controls.



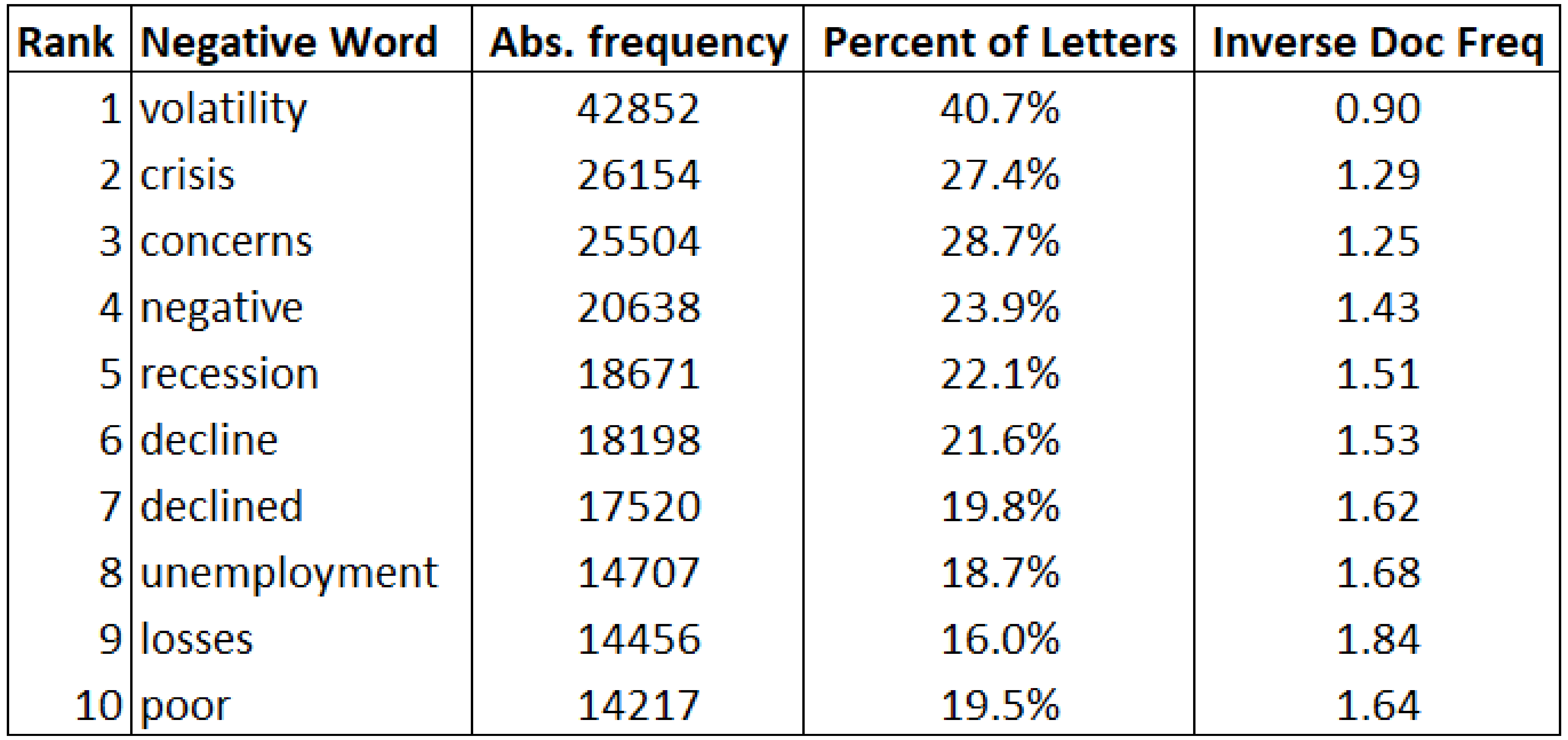
Panel B: Alternative tone measures

Dependent Variable			Flow Filing Month			
	(1)	(2)	(3)	(4)	(5)	(6)
$LMD^{-}_{adj.}$	-0.152***					
	(-2.72)					
Δ LMD^{-}		-0.098**				
		(-2.44)				
LMD^{+}			0.057			
			(1.22)			
LMD^{-}			-0.167***			
			(-3.14)			
$LMD^{+}_{tf.idf}$				0.012**		
				(2.38)		
$LMD^{-}_{tf.idf}$				-0.014***		
				(-3.87)		
HVD^{-}					-0.067**	
					(-2.48)	
$HVD^{-}_{tf.idf}$						-0.006***
						(-2.70)
Controls from Table 3 Column (1)	Y	Y	Y	Y	Y	Y
Flow Reporting Period	N	Y	Y	Y	Y	Y
Lagged Filing Month Flow	N	Y	N	N	N	N
Fund FE	Y	N	Y	Y	Y	Y
Reporting Month FE	Y	Y	Y	Y	Y	Y
Filing Month FE	Y	Y	Y	Y	Y	Y
R^2	0.212	0.104	0.223	0.223	0.222	0.222
Observations	35,359	31,713	38,021	38,021	38,021	38,021

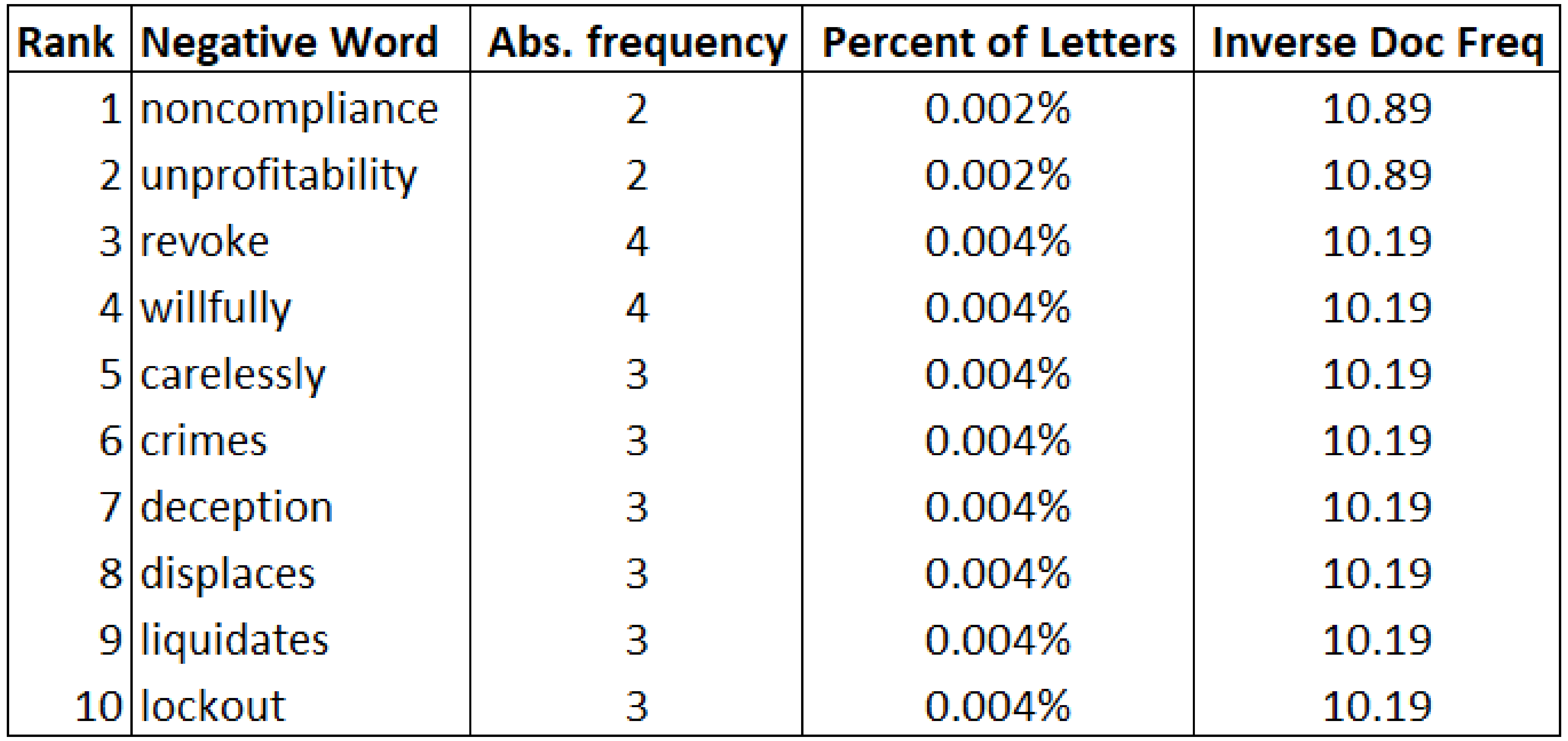
- Column (3): simple word count
  - Only negativity significant.
  - One std. dev. increase in negativity (positivity) associated with -21.5 (5.5) basis points change in flows.
- Column (4): tf.idf weighting as in LM (2011)
  - Positivity and negativity significant.
  - One std. dev. increase in negativity (positivity) associated with -30.0 (15.2) basis points change in flows.

Term weighting

- Top words based on absolute frequency vs. top words based on inverse document frequency (conditional on abs. freq. > 1).



Rank	Negative Word	Abs. frequency	Percent of Letters	Inverse Doc Freq
1	volatility	42852	40.7%	0.90
2	crisis	26154	27.4%	1.29
3	concerns	25504	28.7%	1.25
4	negative	20638	23.9%	1.43
5	recession	18671	22.1%	1.51
6	decline	18198	21.6%	1.53
7	declined	17520	19.8%	1.62
8	unemployment	14707	18.7%	1.68
9	losses	14456	16.0%	1.84
10	poor	14217	19.5%	1.64



Rank	Negative Word	Abs. frequency	Percent of Letters	Inverse Doc Freq
1	noncompliance	2	0.002%	10.89
2	unprofitability	2	0.002%	10.89
3	revoke	4	0.004%	10.19
4	willfully	4	0.004%	10.19
5	carelessly	3	0.004%	10.19
6	crimes	3	0.004%	10.19
7	deception	3	0.004%	10.19
8	displaces	3	0.004%	10.19
9	liquidates	3	0.004%	10.19
10	lockout	3	0.004%	10.19

Source: Hillert et al. (2020): Table IA-3 and unreported results.

- Most frequent words do not necessary indicate bad news (e.g., “unemployment declined”)
- High-idf words sound drastic/significant. → discriminate well between good vs. bad tone.

Recommendations (10)

References

- Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication: introducing a German dictionary. *Journal of Business Economics*, *89*(1), 79-123.
- Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, *20*(2), 639-673.
- Dzieliński, M., Wagner, A. F., & Zeckhauser, R. J. (2018). Straight Talkers, Vague Talkers, and the Value of Firms. Working Paper.
- Garcia, D. (2013). Sentiment during recessions. *Journal of Finance*, *68*(3), 1267-1300.
- Hillert, A., Jacobs, H., & Müller, S. (2014). Media makes momentum. *Review of Financial Studies*, *27*(12), 3467-3501.
- Hillert, A., Niessen-Ruenzi, A., & Ruenzi, S. (2020). Mutual fund shareholder letters: Flows, performance, and managerial behavior. *Performance, and Managerial Behavior.* Working paper.
- Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*(3), 712-729.
- Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307-326.
- Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187-1230.
- Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, *32*(3), 1193-1223.
- Schmeling, M., & Wagner, C. (2019). Does central bank tone move asset prices? Working paper.
- Zhou, D. (2018). Do Numbers Speak Louder Than Words? Working paper.

Agenda

- Main contribution on textual analysis in finance
  - Naïve Bayes approach: Antweiler and Frank (2004)
  - Introduction to machine learning
  - Dictionary approach: Tetlock (2007) and Loughran and McDonald (2011)
- Recommendations for your textual analysis
- Selected topics and papers
  - Readability: Loughran and McDonald (2014, 2020)
  - Textual similarity: Tetlock (2011) and Cohen et al. (2020)

Readability: Loughran and McDonald (2014, 2020)

Readability (1)

How to measure readability?

- Fog Index (see, e.g., Li, 2008)
  - Well established measure for readability.
  - Fog Index = 0.4 x (Words per sentence + Percentage of complex words).
  - A complex word is a word with more than two syllables.
  - Range of the Fog Index
    ≥18: unreadable, 14-18: difficult, 12-14: ideal, 10-12: acceptable, 8-10: childish.

Readability (2)

What is readability?

- No unique definition.
  - *"the ease of understanding or comprehension due to the style of writing" Klare (1963)*
    → sentence length and number of syllables reasonable proxies: Fog Index

  vs.
  - *"the degree to which a given class of people find certain reading matter compelling and comprehensible"* McLuaghlin (1969) and
  - Davison and Kantor (1982, p. 187) point out that the "*background knowledge assumed in the reader"* is more important than *"trying to make a text fit a level of readability defined by a formula."* Loughran and McDonald (2014, p. 1649).
    → Fog Index not an appropriate proxy
- Definition suitable in a business and finance context:
  "we define readability as the ability of individual investors and analysts to assimilate valuation-relevant information from a financial disclosure" Loughran and McDonald (2014, p. 1649).

Readability (3)

Drawbacks of the Fog Index (see, e.g., Loughran and McDonald, 2014)

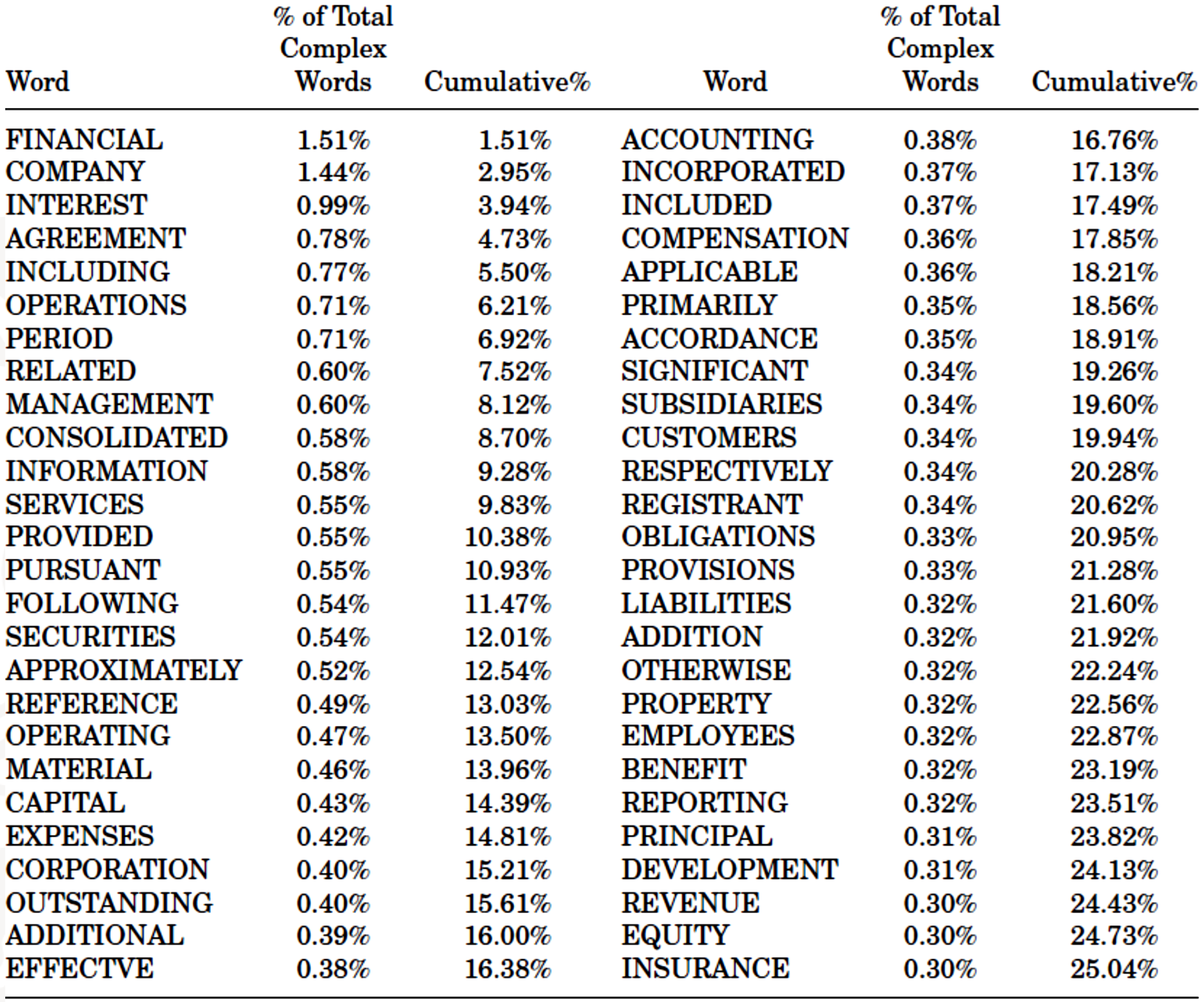
- Business texts contain many words with more than two syllables that are well understood by investors (e.g., company, corporation, telecommunication).
  → one component of the Fog Index is misspecified.
- Furthermore, in financial documents, measuring sentence length is more difficult than in non-financial texts.
  → second component likely to be noisy.

LMD's Alternative to the Fog Index

- Loughran and McDonald (2014) recommend the size of the 10-K complete submission files as readability measure.
- However, Bonsall IV et al. (2017) questions file size as readability proxy, as file size is driven by content unrelated to the underlying text in the 10-K (e.g., HTML, XML, pdf and jpeg file attachments).
- Potential solution: size of the main document / an edited file.

Readability (4)

Most frequent complex words



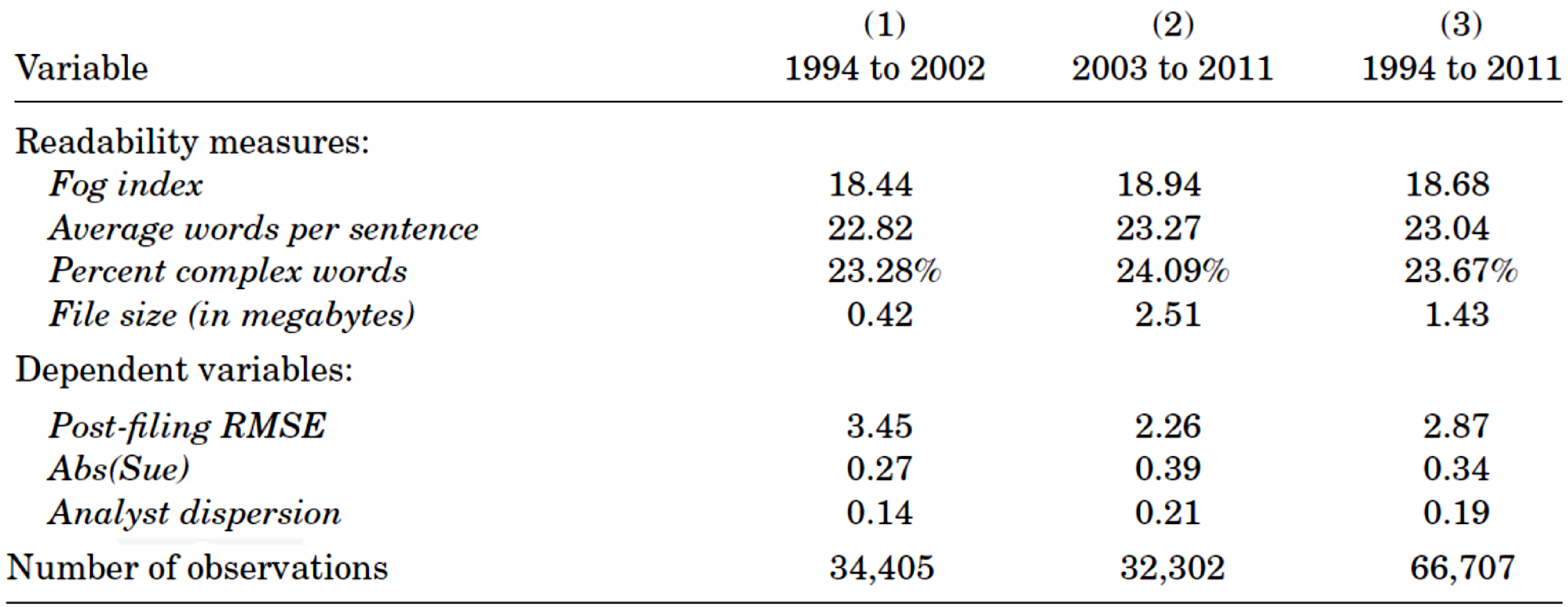
Word	% of Total Complex Words	Cumulative%	Word	% of Total Complex Words	Cumulative%
FINANCIAL	1.51%	1.51%	ACCOUNTING	0.38%	16.76%
COMPANY	1.44%	2.95%	INCORPORATED	0.37%	17.13%
INTEREST	0.99%	3.94%	INCLUDED	0.37%	17.49%
AGREEMENT	0.78%	4.73%	COMPENSATION	0.36%	17.85%
INCLUDING	0.77%	5.50%	APPLICABLE	0.36%	18.21%
OPERATIONS	0.71%	6.21%	PRIMARILY	0.35%	18.56%
PERIOD	0.71%	6.92%	ACCORDANCE	0.35%	18.91%
RELATED	0.60%	7.52%	SIGNIFICANT	0.34%	19.26%
MANAGEMENT	0.60%	8.12%	SUBSIDIARIES	0.34%	19.60%
CONSOLIDATED	0.58%	8.70%	CUSTOMERS	0.34%	19.94%
INFORMATION	0.58%	9.28%	RESPECTIVELY	0.34%	20.28%
SERVICES	0.55%	9.83%	REGISTRANT	0.34%	20.62%
PROVIDED	0.55%	10.38%	OBLIGATIONS	0.33%	20.95%
PURSUANT	0.55%	10.93%	PROVISIONS	0.33%	21.28%
FOLLOWING	0.54%	11.47%	LIABILITIES	0.32%	21.60%
SECURITIES	0.54%	12.01%	ADDITION	0.32%	21.92%
APPROXIMATELY	0.52%	12.54%	OTHERWISE	0.32%	22.24%
REFERENCE	0.49%	13.03%	PROPERTY	0.32%	22.56%
OPERATING	0.47%	13.50%	EMPLOYEES	0.32%	22.87%
MATERIAL	0.46%	13.96%	BENEFIT	0.32%	23.19%
CAPITAL	0.43%	14.39%	REPORTING	0.32%	23.51%
EXPENSES	0.42%	14.81%	PRINCIPAL	0.31%	23.82%
CORPORATION	0.40%	15.21%	DEVELOPMENT	0.31%	24.13%
OUTSTANDING	0.40%	15.61%	REVENUE	0.30%	24.43%
ADDITIONAL	0.39%	16.00%	EQUITY	0.30%	24.73%
EFFECTVE	0.38%	16.38%	INSURANCE	0.30%	25.04%

- The most frequent “complex” words are not complex. They will be easily understood by stock market participants.
- Percentage of complex words is misspecified.

Loughran and McDonald (2014) – Table IV

Loughran and McDonald (2014) (1)

Summary statistics of readability measures



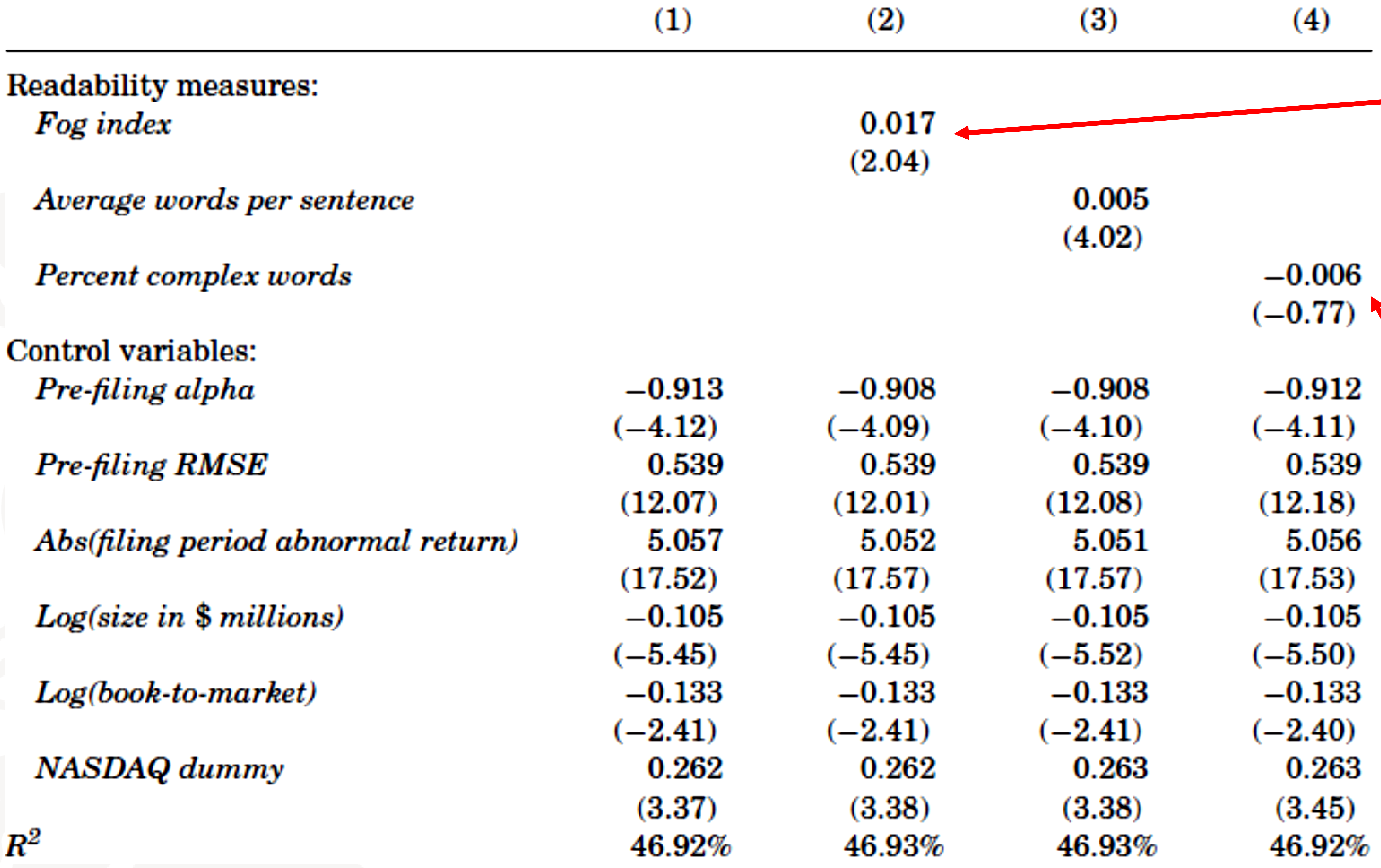
Variable	(1) 1994 to 2002	(2) 2003 to 2011	(3) 1994 to 2011
Readability measures:			
Fog index	18.44	18.94	18.68
Average words per sentence	22.82	23.27	23.04
Percent complex words	23.28%	24.09%	23.67%
File size (in megabytes)	0.42	2.51	1.43
Dependent variables:			
Post-filing RMSE	3.45	2.26	2.87
Abs(Sue)	0.27	0.39	0.34
Analyst dispersion	0.14	0.21	0.19
Number of observations	34,405	32,302	66,707

Loughran and McDonald (2014) –Table II

- “post-filing RMSE”: measures the volatility of the firm’s stock during weeks 2 to 4 after the filing date of the 10-K.
- Idea: if information is difficult to understand there will be a larger delayed reaction by investors.

Loughran and McDonald (2014) (2)

Regressions of post filing volatility on readability



	(1)	(2)	(3)	(4)
Readability measures:				
Fog index		0.017		
		(2.04)		
Average words per sentence			0.005	
			(4.02)	
Percent complex words				−0.006
				(−0.77)
Control variables:				
Pre-filing alpha	−0.913	−0.908	−0.908	−0.912
	(−4.12)	(−4.09)	(−4.10)	(−4.11)
Pre-filing RMSE	0.539	0.539	0.539	0.539
	(12.07)	(12.01)	(12.08)	(12.18)
Abs(filing period abnormal return)	5.057	5.052	5.051	5.056
	(17.52)	(17.57)	(17.57)	(17.53)
Log(size in $ millions)	−0.105	−0.105	−0.105	−0.105
	(−5.45)	(−5.45)	(−5.52)	(−5.50)
Log(book-to-market)	−0.133	−0.133	−0.133	−0.133
	(−2.41)	(−2.41)	(−2.41)	(−2.40)
NASDAQ dummy	0.262	0.262	0.263	0.263
	(3.37)	(3.38)	(3.38)	(3.45)
R^2	46.92%	46.93%	46.93%	46.92%

Loughran and McDonald (2014) – Table III

- Results are in line with the hypothesis that more complex information and/or a more complex presentation of information is associated with higher stock price volatility in the weeks after the filing date.
- As expected, percentage complex words is not related to volatility.
- Regressions include year and industry fixed effects.

Loughran and McDonald (2014) (3)

Alternative measures of readability

- Common words
  - Determine for each word the frequency of filings in which it occurs.
  - Calculate for each filing the average of this proportion for all words in the filing.

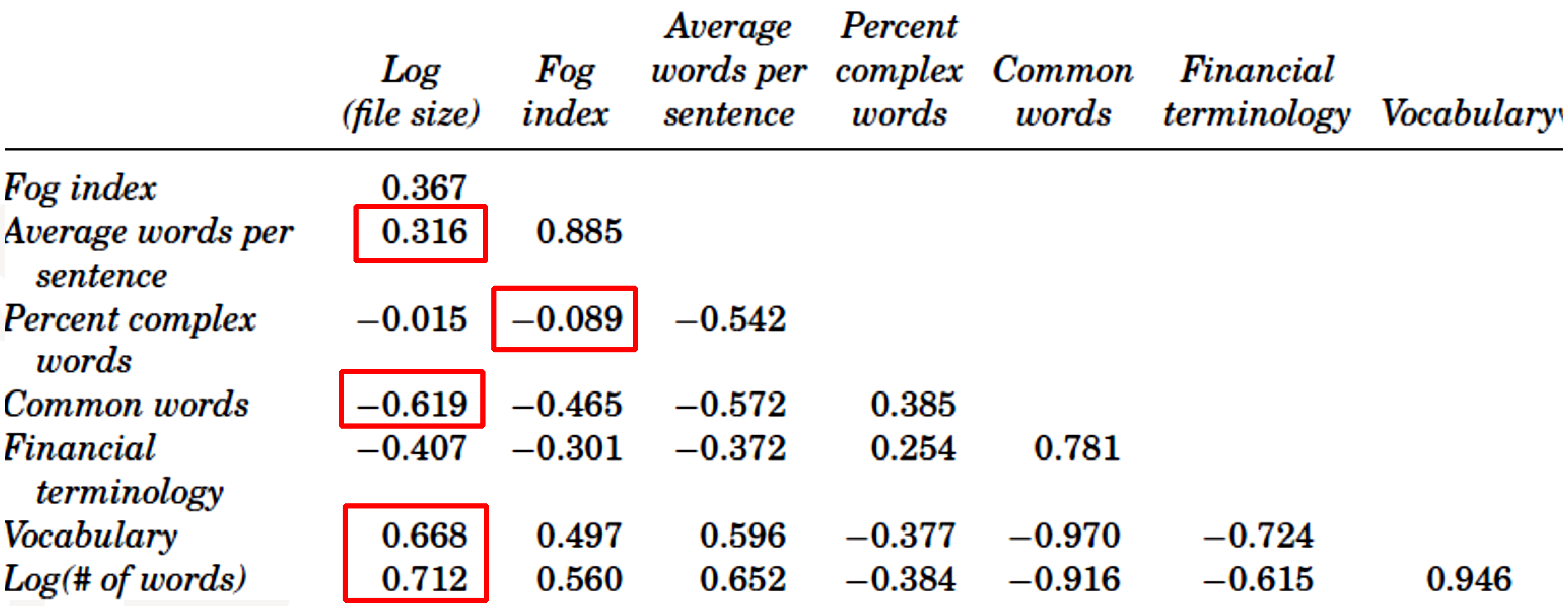
  → idea: the higher the fraction the more ordinary the wording of the 10-K filing.
- Financial terminology
  - Dictionary by Campbell Harvey: http://people.duke.edu/~charvey/Classes/wpg/glossary.htm
  - Number of unique words from Harvey's dictionary divided by total number of unique words.
  - Loughran and McDonald (2014) do not include abbreviations and phrases.

  → idea: higher fraction indicates more value relevant information.
- Vocabulary
  - Number of unique words in a filing divided by total number of words in master dictionary.

  → idea: extensive vocabulary makes document less comprehensible.
- Number of words
  - Longer document → less readable

Loughran and McDonald (2014) (4)

Correlations of different readability measures



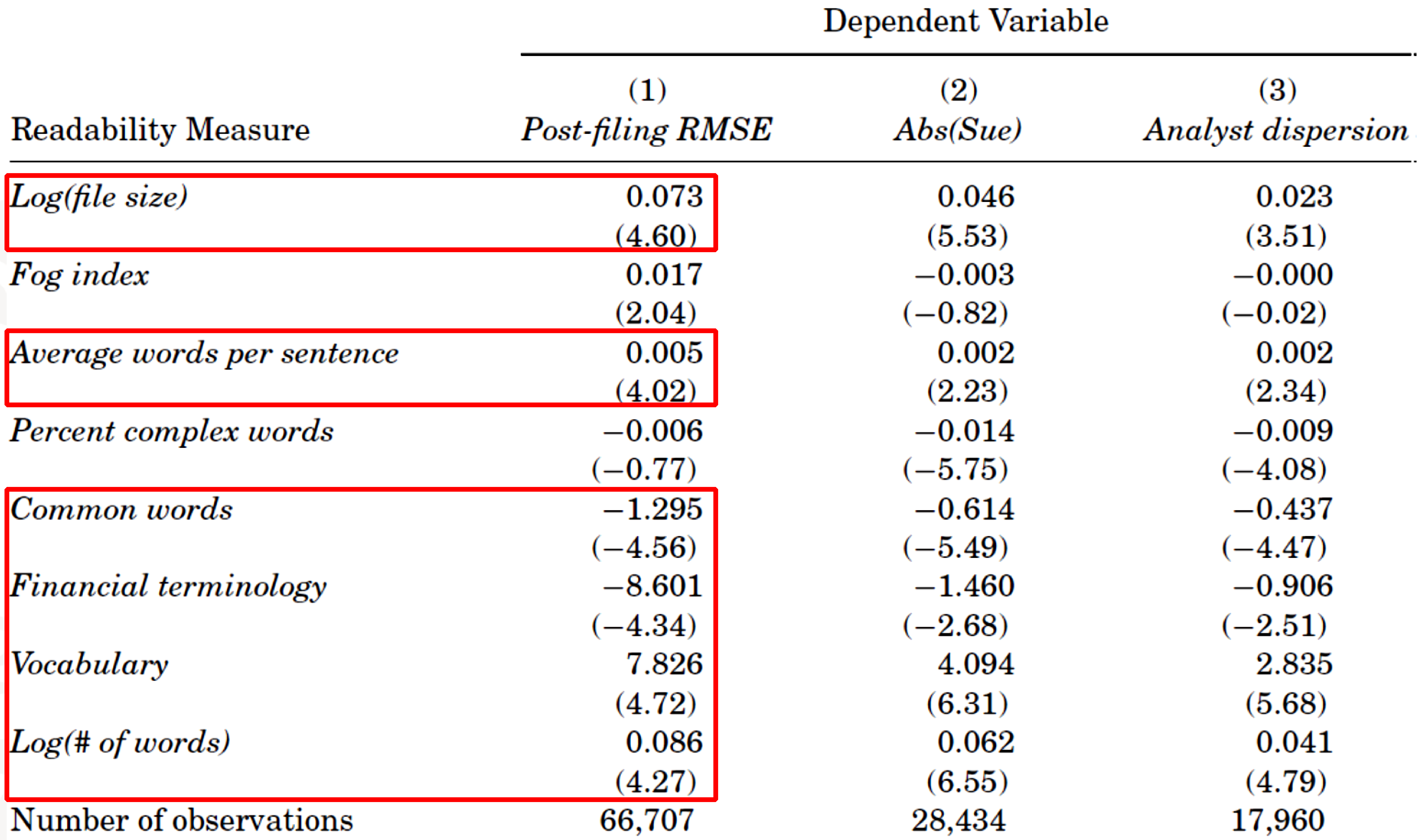
	Log (file size)	*Fog index*	*Average words per sentence*	*Percent complex words*	*Common words*	*Financial terminology*	*Vocabulary*
Fog index	0.367						
Average words per sentence	0.316	0.885					
Percent complex words	−0.015	−0.089	−0.542				
Common words	−0.619	−0.465	−0.572	0.385			
Financial terminology	−0.407	−0.301	−0.372	0.254	0.781		
Vocabulary	0.668	0.497	0.596	−0.377	−0.970	−0.724	
Log(# of words)	0.712	0.560	0.652	−0.384	−0.916	−0.615	0.946

Loughran and McDonald (2014) – Table IV

- Correlations in line with motivation on previous slide
  - ○ Positive relation between file size and (1) average WPS, (2) vocabulary, and (3) # of words.
  - ○ Negative relation between file size and (1) common words and (2) terminology.
- Surprising that the percentage of complex words is negatively related to the Fog index.

Loughran and McDonald (2014) (5)

Regressions of uncertainty on readability measures



	Dependent Variable		
	(1)	(2)	(3)
Readability Measure	*Post-filing RMSE*	*Abs(Sue)*	*Analyst dispersion*
Log(file size)	0.073	0.046	0.023
	(4.60)	(5.53)	(3.51)
Fog index	0.017	−0.003	−0.000
	(2.04)	(−0.82)	(−0.02)
Average words per sentence	0.005	0.002	0.002
	(4.02)	(2.23)	(2.34)
Percent complex words	−0.006	−0.014	−0.009
	(−0.77)	(−5.75)	(−4.08)
Common words	−1.295	−0.614	−0.437
	(−4.56)	(−5.49)	(−4.47)
Financial terminology	−8.601	−1.460	−0.906
	(−4.34)	(−2.68)	(−2.51)
Vocabulary	7.826	4.094	2.835
	(4.72)	(6.31)	(5.68)
Log(# of words)	0.086	0.062	0.041
	(4.27)	(6.55)	(4.79)
Number of observations	66,707	28,434	17,960

Loughran and McDonald (2014) – Table IX

Econometric setup

- These are 24 separate regressions!
- Regressions include controls, year and industry FEs.

Result

- Larger filings, and filings with more (unique) words,
- filings with longer sentences,
- and fewer financial terms

have higher post-announcement volatility.

Summary on readability (1)

(Critical) discussion about readability

- Do investors really read filings?
  - Loughran and McDonald (2017) find little interest in form 10-K filings.
    - Robot vs. non-robot requests: more than 50 requests per day from an IP address → robot.
    - Average number of downloads of 10-Ks on the filing + following day = 28.4
  - Cohen et al. (2020; see next section) find strong return predictability based on text changes in 10-Ks/10-Qs.

- Kim et al. (2019) suggest a fix to the Fog index.
  - Idea: main problem is the “complex” words → identify and exclude “complex” words that are standard business language.
  - Create list of 2,028 words with >2 syllables that are not complex, e.g., “auditor”, “acquisition”.
    → not counted in the computation of the index.
  - Average modified Fog index vs. Fog index: 12.96 vs. 19.69.

Summary on readability (2)

(Critical) discussion about readability - continued

- General problems of readability scores
  - Fog index and similar indices (e.g., Flesch-Kincaid index) are primarily used for grade-textbooks.
  - Industry component of readability (e.g., name of chemicals).
  - Attachments (e.g., legal contracts) affect readability scores.

- Readability vs. firm complexity
  - Readability may just reflect the complexity of firms' businesses.
    Example: McDonalds vs. 3M (> 60,000 products; adhesives, passive fire protection, PPE, dental products, electronic circuits, etc.)
  - Loughran and McDonald (2020) develop a complexity dictionary
    - 374 complex words: acquire, lease, contract, subsidiary.
    - Complexity predicts higher audit fees.
    - Unfortunately, they do not relate readability measures to their complexity measure.

Textual similarity: Tetlock (2011) and Cohen et al. (2020)

Textual similarity – Introduction

Why is textual similarity interesting?

- Distinguish new and old/stale information
- Example: Huberman and Regev (2001) show that a reprinted newspaper article (→ no new information) leads to a large stock price reaction (May 4).

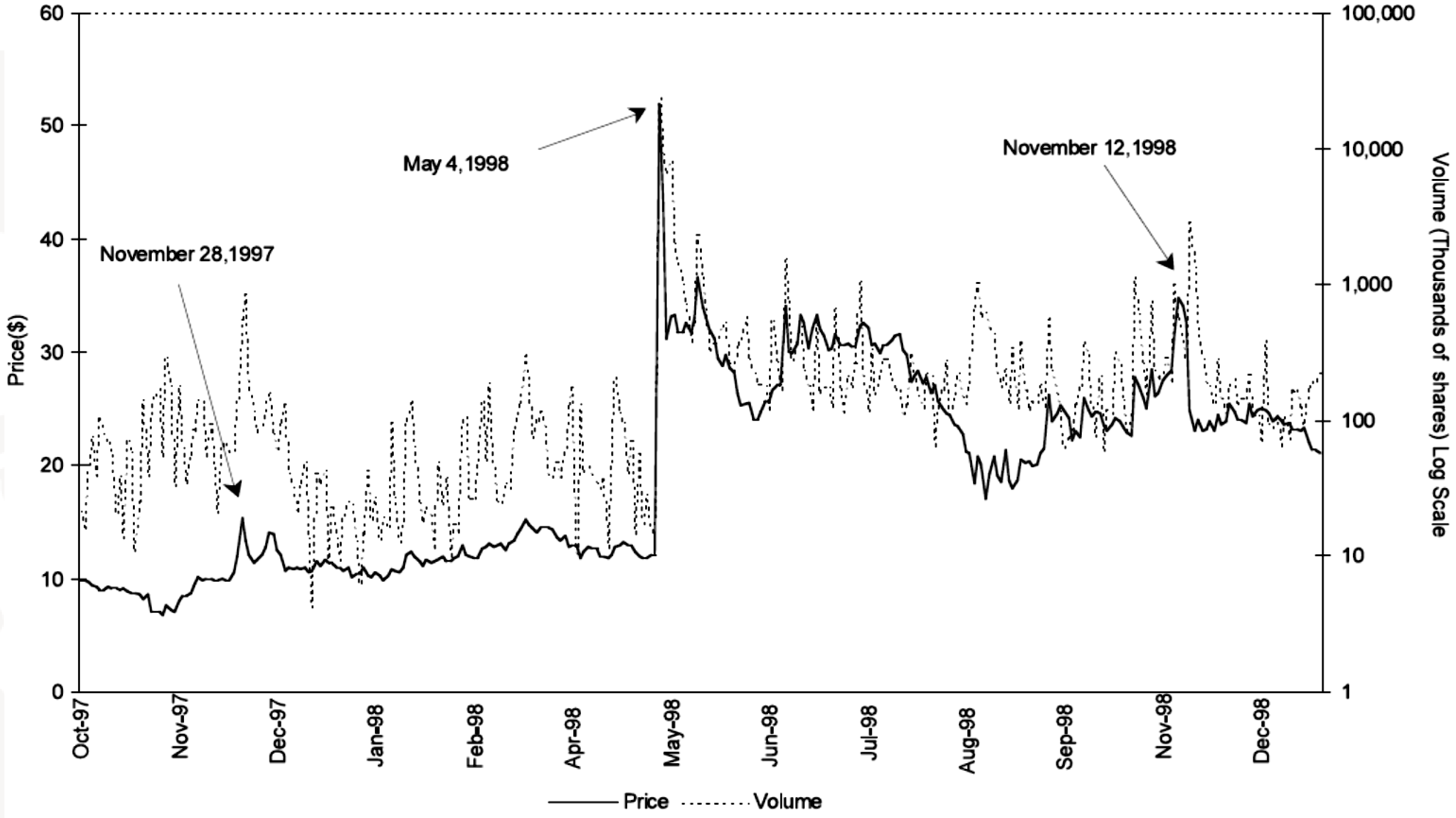


Figure 1. ENMD closing prices and trading volume, October 1, 1997, to December 30, 1998.

- More generally: market efficiency (stock prices reflect all available information).

Tetlock (2011) (1)

Summary of Tetlock (2011)

- Sample: firm-specific news stories in the DJ Newswire from Nov 1996 to Oct 2008.
- Initial and delayed market reaction to new and old/stale news.
- Staleness: average similarity to the previous ten news stories.
- Average market reaction to new news (bottom staleness decile) is 413 bp (75 bp) stronger than the reaction to stale news (top staleness decile) for equal (value) weighted portfolios.
  → investors differentiate between new and old news.
- However, 26 bp stronger return reversal in the week after stale news.
  → investors' initial response not perfect.

→ **How does Tetlock measure similarity?**

Jaccard Similarity

The Jaccard similarity measures the similarity/diversity of two sets

- $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$
- “Intersection over Union”
- Example:
  - A: “This is the first text of the Jaccard example.”
  - B: “This sentence represents the second Jaccard example.”

  → $A \cap B$: “this”, “the”, “Jaccard”, “example” → $|A \cap B| = 4$

  → $A \cup B$: “this”, “is”, “the”, “first”, “text”, “of”, “Jaccard”, “example”, “sentence”, “represents”, “second”. → $|A \cup B| =$11

  → Jaccard similarity = 4 / 11=0.3636

Tetlock (2011) (2)

Additional editing of texts

Tetlock (2011) footnote 2:

- “Before identifying unique words and bigrams, I exclude a standard list of 119 extremely common words such as “into,” “so,” and “that”; 42 common numbers (0 through 9 and 1978 through 2009); and 27 terms that are ubiquitous in financial news stories, such as “Dow Jones,” “New York,” and “newswire.””
- “I also use a standard word-stemming algorithm to equate all similar forms of a word—e.g., “changing” and “changed” are both derivatives of “change.””

→We will compute textual similarity for a sample of 10-K filings (Problem 14).

→Why does it economically make sense to analyze 10-K similarity? →Cohen et al. (2020)

Cohen et al. (2020): Motivation and research question of the study

- Idea: when confronted with repetitive tasks, people tend to use the same approach as last time (“copy and paste”).
- This also hold for repetitive tasks in corporations like preparing annual reports. Example:
  - Beginning of Apple’s 2018 form 10-K filing (“Item 1. Business”)
    “The Company designs, manufactures and markets mobile communication and media devices and personal computers, and sells a variety of related software, services, accessories and third-party digital content and applications. The Company’s products and services include iPhone®, iPad®, Mac®, Apple Watch®, AirPods®, Apple TV®, HomePod™, a portfolio of consumer and professional software applications, iOS, macOS®, watchOS® and tvOS™ operating systems, iCloud®, Apple Pay® and a variety of other accessory, service and support offerings.”
    Available at: https://www.sec.gov/Archives/edgar/data/320193/000032019318000145/a10-k20189292018.htm
  - Beginning of Apple’s 2017 form 10-K filing (“Item 1. Business”)
    “The Company designs, manufactures and markets mobile communication and media devices and personal computers, and sells a variety of related software, services, accessories, networking solutions and third-party digital content and applications. The Company’s products and services include iPhone®, iPad®, Mac®, Apple Watch®, Apple TV®, a portfolio of consumer and professional software applications, iOS, macOS®, watchOS® and tvOS™ operating systems, iCloud®, Apple Pay® and a variety of accessory, service and support offerings.”
    Available at: https://www.sec.gov/Archives/edgar/data/320193/000032019317000070/a10-k20179302017.htm

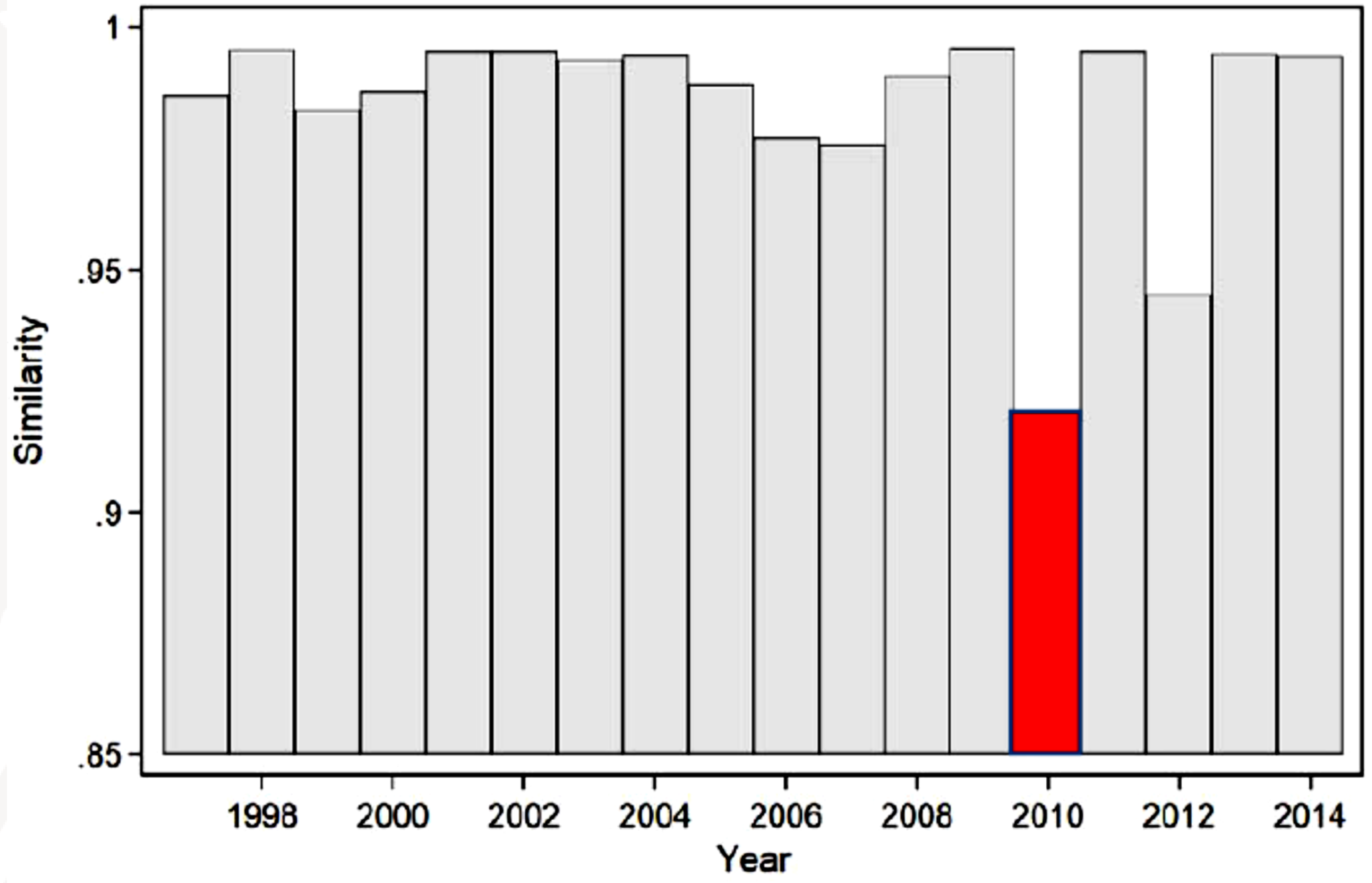
Cohen et al. (2020): Motivation and research question of the study

- This “copy-and-paste” approach implies that a drop in similarity of a company’s annual reports indicates that some major change is happening at the firm.
- Knowing about the change early is potentially valuable.
  - Positive news: buy the stock.
  - Negative news: (short) sell the stock.
  - information on textual similarity not readily available to all investors?

Cohen et al. (2020) – “Lazy Prices”; The Value of Text-based Information (3)

Similarity of annual reports

- What is the average similarity of annual reports? Are drops in similarity informative?
- Example: Baxter International Inc.
  - Large U.S. pharma firm producing among other things products to intravenously deliver fluids and drugs to patients.
  - Similarity of Baxter’s 10-Ks from 1997 to 2014



- Annual reports very similar with Jaccard similarities ranging from 0.97 to 1.
- Sharp drop in similarity in 2010 → what is the reason?
- Source: Figure 2 of Cohen et al. (2018, working paper version)

Example: Baxter Inc. – Events in 2010

What happened at Baxter in 2010?

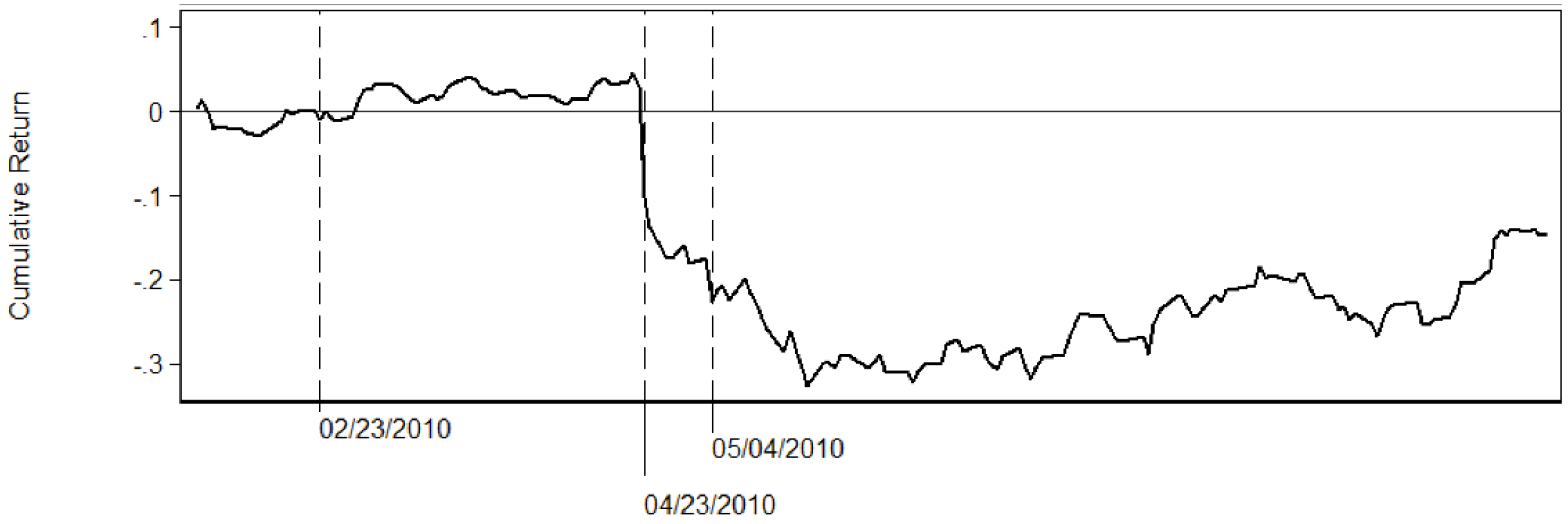
- February 23, 2010: Baxter filed its fiscal year 2009 form 10-K with the Securities and Exchange commission (SEC) → financial report is publicly available.
- April 23, 2010: the New York Times (https://www.nytimes.com/2010/04/24/business/24pump.html) reports that the Food and Drug Administration (FDA) will tighten its regulation on medical equipment. More precisely, drug companies need to provide more test data before infusion pumps get approved.
- May 4, 2010: the New York Times (https://www.nytimes.com/2010/05/04/business/04baxter.html) writes that Baxter will recall infusion pumps under an agreement with regulators and that Baxter agrees to pay a charge between $400 and $600 million.

→Negative news for Baxter and is associated lower firm value.

→When do investors price this news into Baxter’s stock price?

Example: Baxter Inc. – Stock price reaction in 2010

Cumulative return of Baxter's stock around the three events



Source: Figure 3 of Cohen et al. (2020)

- No reaction to Baxter's form 10-K filing.
- Strong stock price decline in response to FDA's tighter regulation.
- After Baxter's agreement with regulators follows a further drop in the Baxter's value.

→ Was there already information in Baxter's 10-K filing? If so, investors could have made significant profits.

Cohen et al. (2020) – “Lazy Prices”; The Value of Text-based Information (6)

Example: Baxter Inc. – New information in 2010’s 10-K

What text parts have been added to the 2010 form 10-K relative to the 2009 form 10-K:

- Two examples (text changes are underlined):
  - In Item 1A. Risk Factors: “It is possible that substantial additional charges, including significant asset impairments, related to COLLEAGUE may be required in future periods, based on new information, changes in estimates, and modifications to the current remediation plan.”
  - In the “Notes to Consolidated Financial Statements”: “In 2009, the company recorded a charge of $27 million related to planned retirement costs associated with SYNDEO and additional costs related to the COLLEAGUE infusion pump. This charge consisted of $14 million for cash costs and $13 million related to asset impairments. The reserve for cash costs primarily related to customer accommodations and additional warranty costs.”

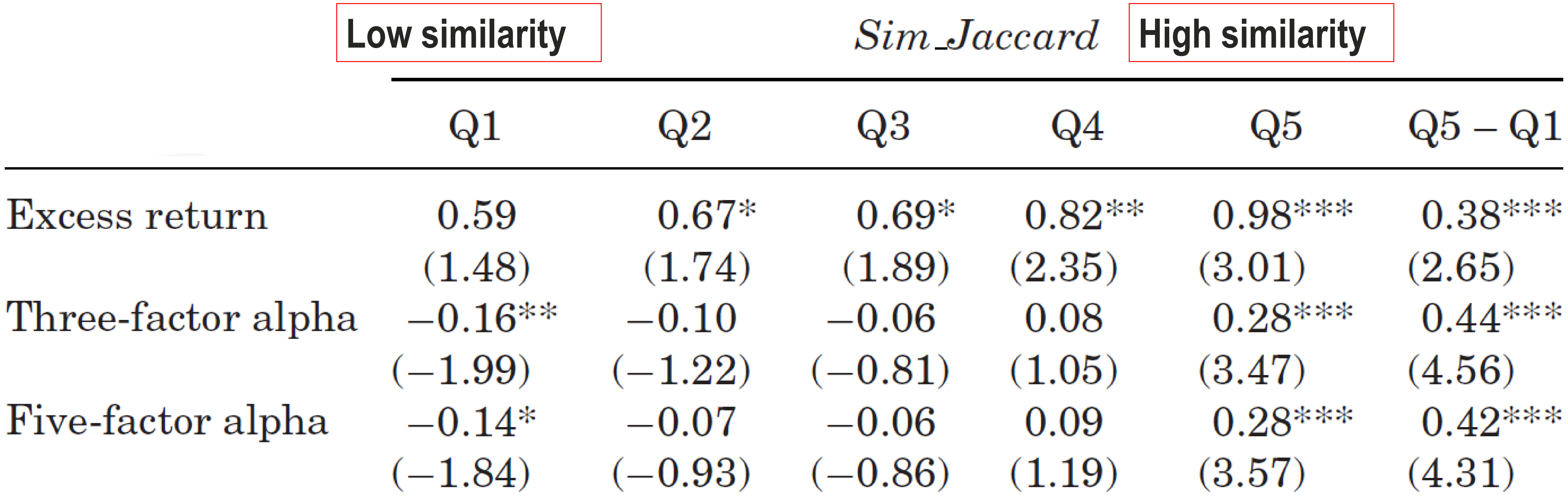
  The 10-K is available at: https://www.sec.gov/Archives/edgar/data/10456/000095012310015380/c54958e10vk.htm.

→There are clear hints that Baxter faces problems with its infusion pumps.

→Could have been very profitable for investors to carefully read (and identify changes in) Baxter’s 10-K.

Cohen et al. (2020) – “Lazy Prices”; The Value of Text-based Information (7)

Empirical results of a textual similarity investment strategy



	Low similarity		*Sim_Jaccard*		High similarity	
	Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess return	0.59	0.67*	0.69*	0.82**	0.98***	0.38***
	(1.48)	(1.74)	(1.89)	(2.35)	(3.01)	(2.65)
Three-factor alpha	−0.16**	−0.10	−0.06	0.08	0.28***	0.44***
	(−1.99)	(−1.22)	(−0.81)	(1.05)	(3.47)	(4.56)
Five-factor alpha	−0.14*	−0.07	−0.06	0.09	0.28***	0.42***
	(−1.84)	(−0.93)	(−0.86)	(1.19)	(3.57)	(4.31)

- First row shows the average monthly portfolio return minus the risk-free rate (“excess return”).
- Second (third) row reports the Fama and French 3-factor (5-factor) alphas.
- Numbers in parenthesis show t-statistics.
- *, **, *** indicate significance at the 10%, 5%, and 1% level, respectively.

Source: Table II of Cohen et al. (2020)

- Significant return difference of 0.38% per month between the portfolio of stocks with high similarity and the one with low similarity.
- This return difference cannot be explained by systematic risk factors.

Summary of the chapter

What have we learnt in this part?

- Textual analysis is a growing research field in accounting and finance.
- Besides analyzing document tone/sentiment there are further dimensions that are economically relevant including:
  - Readability
  - Staleness of information
  - Precision of information: LMD uncertainty and modal weak
  - Forward vs. backward looking information
- Simply applying a dictionary to a new type of document is unlikely to result in a top contribution, but quantifying verbal information to empirically testing so far untested ideas seems promising.

Thank you very much for your attention!