Word Power: A new approach for content analysis

Narasimhan Jegadeesh
Emory University
Andrew Di Wu
University of Pennsylvania

Research Questions

- Does qualitative text convey information about stock value beyond quantitative data?
- Do investors efficiently incorporate qualitative information into prices?

Literature Review

- Bag-of-Words Approach
  - Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008)
    - Harvard Psychosociological Dictionary
    - Media accounts
  - Li (2006)
    - “risk” (“risk”, “risks”, and “risky”)
    - “uncertainty” (“uncertain”,“uncertainty”, and “uncertainties”)
    - MD&A section of 10-Ks
  - Loughran and McDonald (2011)
    - Compile a lexicon of negative and positive words from 10-Ks
  - Feldman, Govindaraj, Livnat, and Segal (2010)
    - LM Dictionary, MD&A of 10-Ks and 10-Qs

Related approaches for content analysis

- Algorithmic approach
  - Das and Chen (2007)
- Start with a “training sample” with text classified as optimistic, neutral and pessimistic. Uses different algorithms to detect the words that best discriminate among these categories.
- Our approach uses contemporaneous returns to calibrate negative/positive document tone

Content Analysis

- Lexicon
  - Harvard List
  - LM list
- Term Weighting
  - Unweighted: All words in the lexicon have the same impact
  - idf: Word weights inversely proportional to the frequency of occurrence in the sample of documents

Our Focus and Main Results

- Term weights based on past market reactions
  - Low correlation with document tone scores with other weighting schemes even with the same underlying lexicon
  - More accurate document tone score for both positive and negative words
  - Choice of term weights at least as important as choice of word lexicon

Data

- First filing of 10-Ks for the year from EDGAR
- Non-Financials
- Minimum price of $3 per share on the filing date

Parsing the 10-K

- Exclude tables and exhibits
- Include only words in the dictionary from 2of12inf dictionary

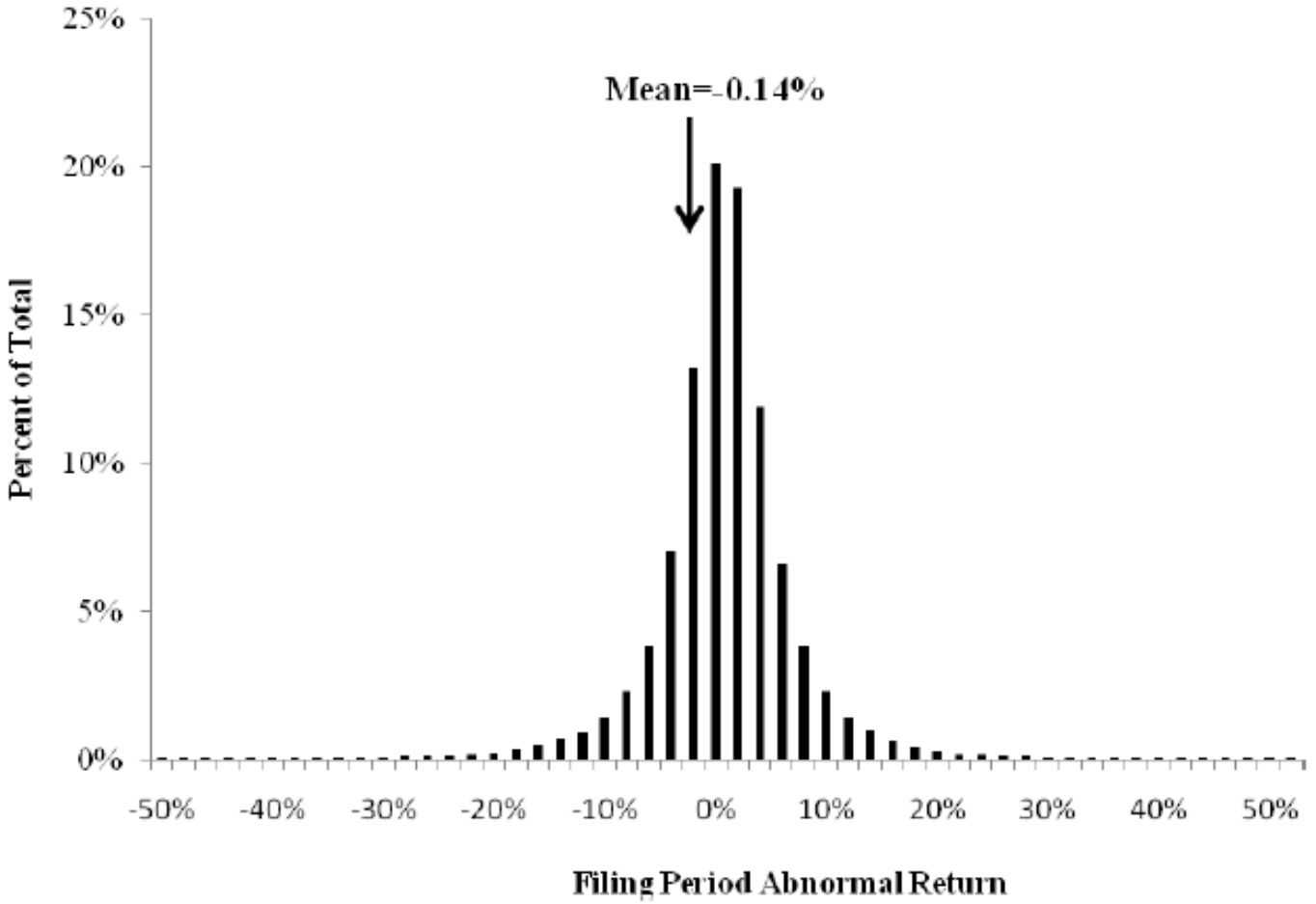
(wordlist.sourceforge.net/12dicts-readme.html)

- Exclude common stop words and single character words

Year	# of Firms	Size ($bln)	
		Mean	Median
1995	1,429	$2.03	$0.34
1996	2,330	$1.82	$0.22
1997	3,607	$1.65	$0.20
1998	3,619	$2.05	$0.23
1999	3,337	$2.79	$0.23
2000	3,533	$3.21	$0.32
2001	3,066	$3.20	$0.33
2002	2,850	$3.07	$0.36
2003	2,629	$2.70	$0.33
2004	3,013	$3.22	$0.45
2005	2,940	$3.46	$0.50
2006	2,904	$3.87	$0.59
2007	2,845	$4.38	$0.65
2008	2,687	$4.34	$0.58
1995-2008	40,789	$3.02	$0.36

Figure 1
Distribution of Filing Period Abnormal Returns

This figure plots that distribution of filing period abnormal return, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date + 3]. Our sample is comprised of 40,789 unique 10-Ks from 1995 to 2008.



Lexicon

- LM positive and negative word lists
  - 353 positive words and 2337 negative words including all inflections
  - We manually assign each inflection to root words
    - E.g. falsify includes falsified, falsifies, falsification, falsifications, and falsifying
    - Defend and defendant are different root words
- The inflection-adjusted lexicon has 122 and 716 positive and negative root words

Inverse document frequency weights

$$w_j^{idf} = \log \frac{N}{df_j}$$

N: Number of documents in the sample
df_j: Number of documents in which word *j* occurs

Word Power Weight - Methodology

$$Score_i = \sum_{j=1}^{J} \left(w_j F_{i,j} \right) \times \frac{1}{a_i}, \qquad (4)$$

W_j: Weight for word j
$F_{i,j}$: Number of occurrences of word j in document i
a_i: Total number of words in Document i
J: Total number of words in the positive/negative word list

Document tone and returns

$$r_i = a + b \times \left(\sum_{j=1}^{J} \left(w_j F_{i,j} \right) \times \frac{1}{a_i} \right) + \varepsilon_i$$
$$= a + \left(\sum_{j=1}^{J} \left(\left[b \times w_j \right] F_{i,j} \right) \times \frac{1}{a_i} \right) + \varepsilon_i, \quad (5)$$

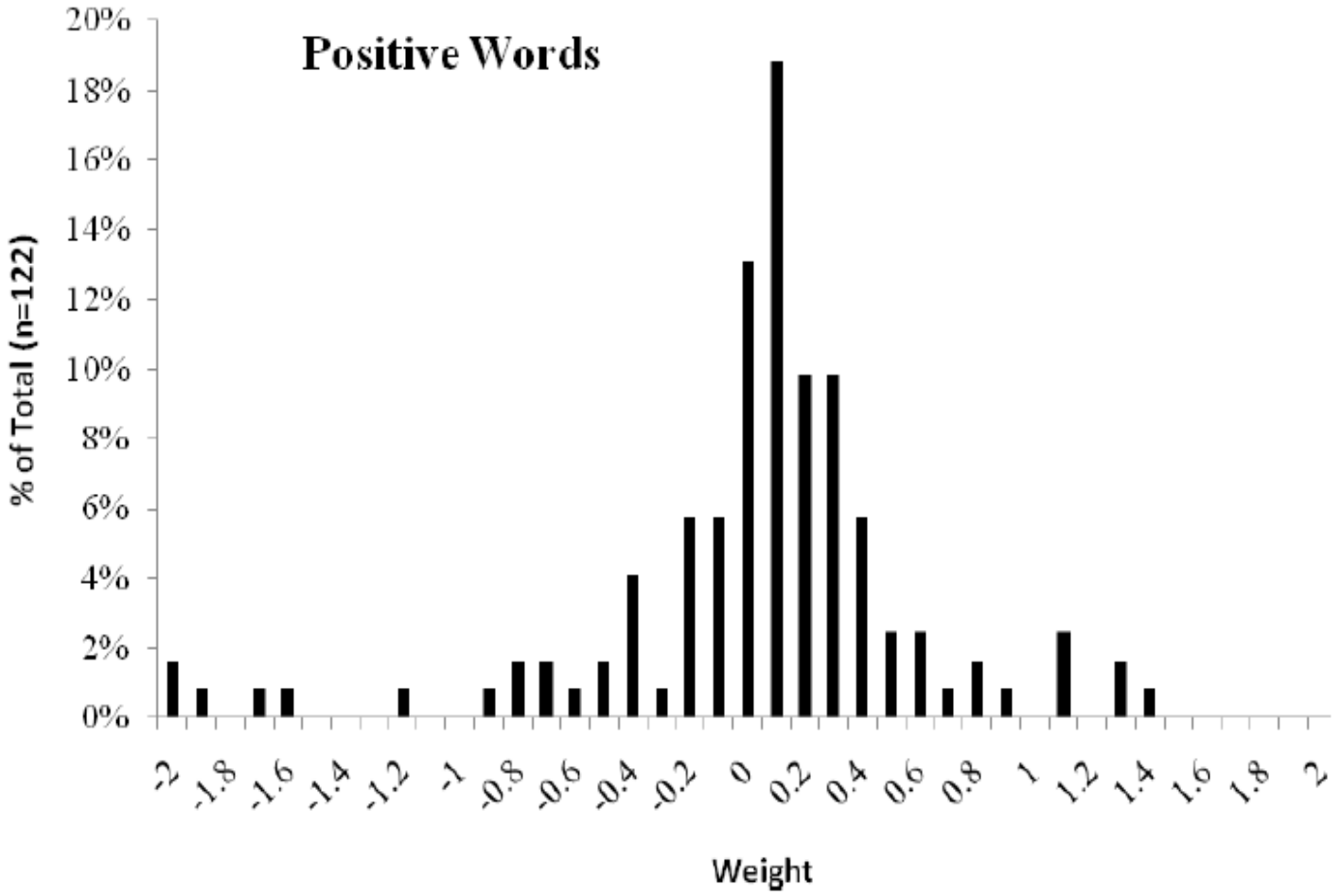
where r_i is the abnormal return when the i^{th} document is released.

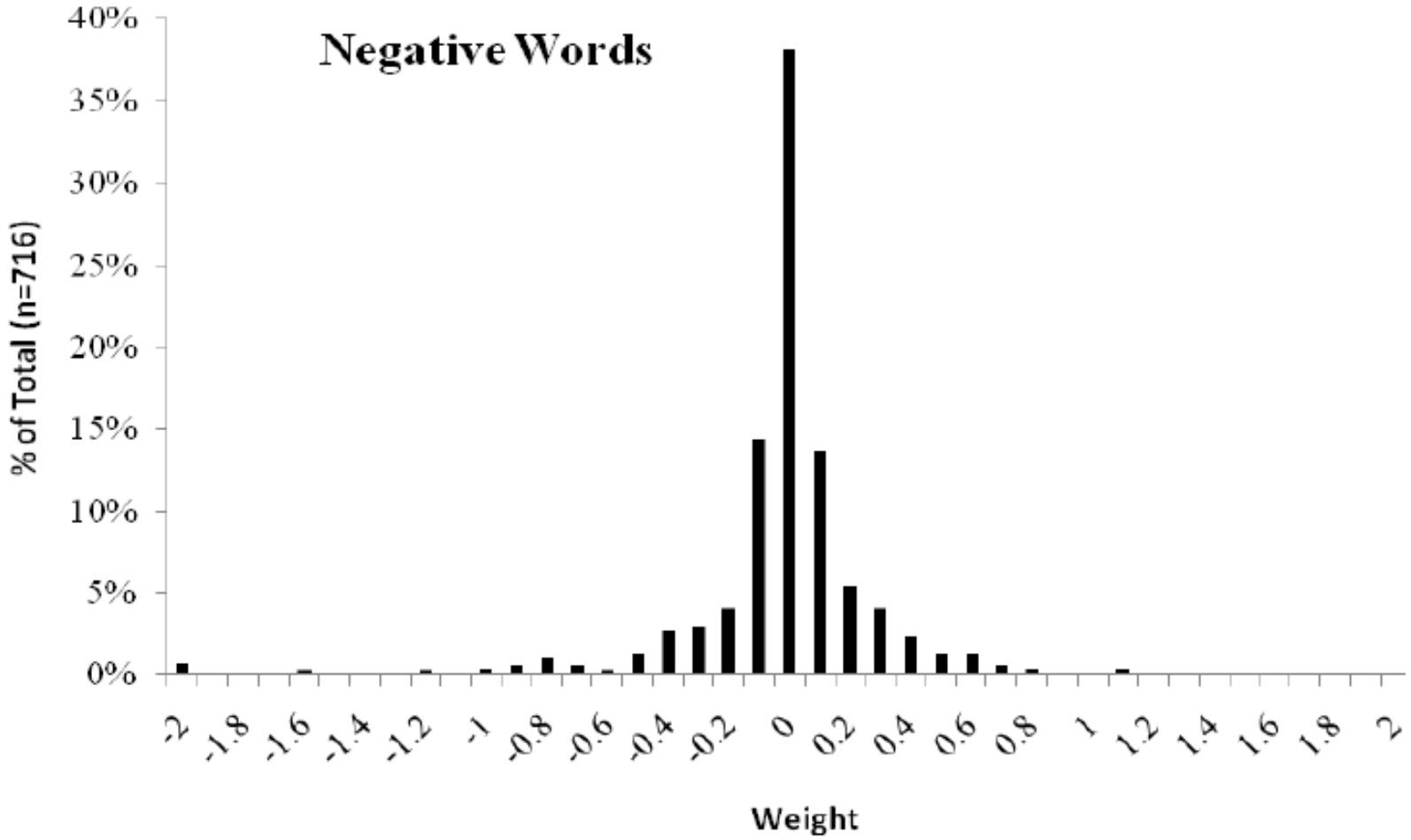
Estimate WP weights

$$r_i = a + \left(\sum_{j=1}^{J} \left(B_j F_{i,j} \right) \times \frac{1}{a_i} \right) + \varepsilon_i,$$

$$b \times w_j$$

$$\hat{w}_j = \frac{\hat{B}_j - \bar{B}}{\text{Standard Deviation } \left(\hat{B}_j \right)},$$


Positive Words
% of Total (n=122)
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
-2
-1.8
-1.6
-1.4
-1.2
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Weight


Negative Words
% of Total (n=716)
40%
35%
30%
25%
20%
15%
10%
5%
0%
-2
-1.8
-1.6
-1.4
-1.2
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Weight

Inverse document frequency weights

$$w_j^{idf} = \log \frac{N}{df_j}$$

N: Number of documents in the sample
df_j: Number of documents in which word *j* occurs

Panel A: Positive Words

Weight Quintile	Frequency Quintile (%) 1	2	3	4	5	Row Total
1	40	36	16	8	0	25
2	8.33	4.17	12.5	25	50	24
3	8	8	12	36	36	25
4	12.5	16.67	29.17	29.17	12.5	24
5	33.33	33.33	33.33	0	0	24

Panel B: Negative Words

Weight Quintile	Frequency Quintile (%) 1	2	3	4	5	Row Total
1	46.53	31.25	15.97	4.86	1.39	144
2	6.99	15.38	19.58	30.77	27.27	143
3	0.7	5.59	16.78	27.27	49.65	143
4	7.69	12.59	26.57	32.17	20.98	143
5	38.46	34.97	20.98	4.9	0.7	143

Correlation Between WP and *idf* weights and document scores

	Word List	10-K
Negative Words	-0.045	0.031
Positive Words	0.143	-0.295

Low correlation between document scores assigned by tf.idf weights and WP weights.

Inverse Document Frequency Score

$$w_{i,j}^{tf.idf} = 1 + \log\left(tf_{i,j}\right) w_j^{idf}, \text{ if } \quad tf_{i,j} > 0$$
$$= 0 \qquad\qquad \text{otherwise} \tag{2}$$

where $tf_{i,j}$ is the frequency of occurrence of the word j in document i. The document score using the idf word weights, which we refer to as $Score_i^{tf.idf}$ or $tf.idf$ score, is computed as:

$$Score_i^{tf.idf} = \frac{1}{\left(1 + \log a_i\right)} \sum_{j=1}^{J} w_{i,j}^{tf.idf}, \tag{3}$$

Panel A: Positive Words

Most Impactful Words			Least Impactful Words		
	WP Rank	*idf* Rank		WP Rank	*idf* Rank
ingenuity	1	14	lucrative	122	13
acclaimed	2	7	tremendous	121	35
influential	3	26	receptive	120	30
regain	4	39	happy	119	9
enthusiasm	5	29	beautiful	118	15
optimistic	6	42	conducive	117	27
revolutionize	7	19	smoothes	116	60
courteous	8	20	vibrant	115	16
incredible	9	3	outperformed	114	32
excited	10	48	transparent	113	43

Panel B: Negative Words

Most Impactful Words			Least Impactful Words		
	WP Rank	*idf* Rank		WP Rank	*idf* Rank
imperil	1	18	disorderly	716	3
insubordination	2	20	ridicule	715	2
vitiate	3	38	disgrace	714	1
bailout	4	31	derogatory	713	4
unwelcome	5	5	immoral	712	23
dismal	6	10	disassociate	711	35
denigrate	7	36	mischief	710	27
inadvisable	8	56	extenuating	709	34
turbulent	9	140	dispossess	708	8
undocumented	10	55	irreconcilable	707	11

Determinants of Document Tone - Independent Variables

- Size: Natural logarithm of the market capitalization of equity at the end of month before the 10-K filing date.
- BM: The ratio of the book value of equity as of the fiscal year end in the 10-K.
- Volatility: The standard deviation of the firm-specific component of returns estimated using up to 60 months of data as of the end of the month before the filing date. We estimate volatility for all firms with at least 12 months of data during this 60-month period.
- Turnover: Natural logarithm of the number of shares traded during the period from six to 252 trading days before the filing date divided by the number of shares outstanding on the filing date.

Determinants of Document Tone - Independent Variables

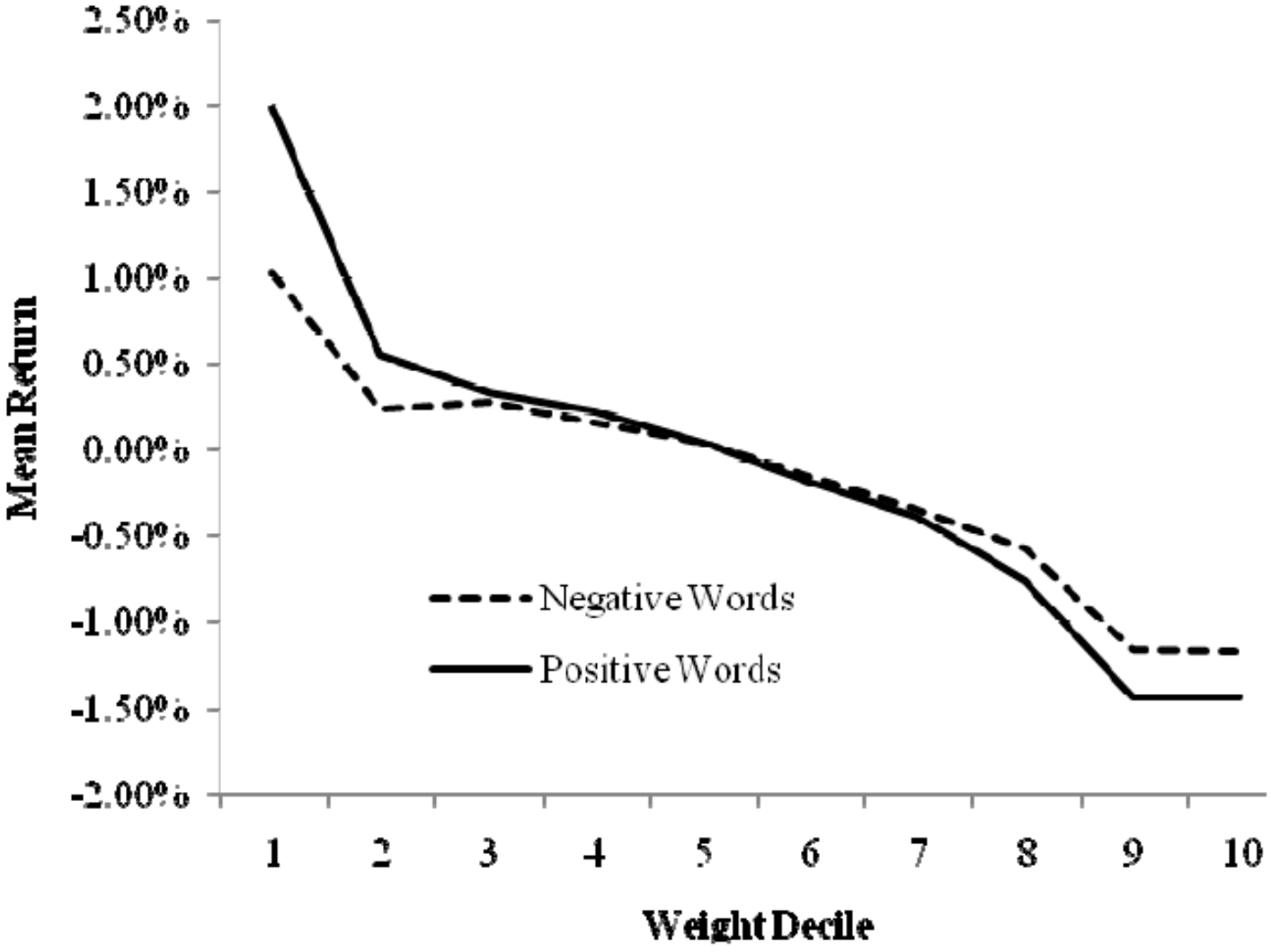
- EAD-Ret: The cumulative return over the three-day window [t-1, t+1] around the latest earnings announcement date minus the contemporaneous CRSP value-weight index return over the same period.
- Accruals: We compute accruals as in Sloan (1996). Specifically, accruals is one-year change in current assets excluding cash minus change in current liabilities excluding long-term debt in current liabilities and taxes payables minus depreciation divided by average total assets.

Word Power Weight - Methodology

$$Score_i = \sum_{j=1}^{J} \left(w_j F_{i,j} \right) \times \frac{1}{a_i}, \qquad (4)$$

W_j: Weight for word j
$F_{i,j}$: Number of occurrences of word j in document i
a_i: Total number of words in Document i
J: Total number of words in the positive/negative word list

	Negative Tone $Score_i$	Positive Tone $Score_i$
Independent Variables		
Size	0.024	-0.001
	(8.36)	(-0.44)
BM	0.043	0.031
	(2.77)	(2.95)
Volatility	-0.192	-0.278
	(-3.69)	(-4.04)
Turnover	-0.022	-0.028
	(-2.78)	(-2.97)
EAD-Ret	0.017	0.015
	(0.45)	(0.11)
Accruals	-0.044	-0.076
	(-0.32)	(-0.08)
$Score_{i-1}$	0.521	0.741
	(6.99)	(7.57)


2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
-0.50%
-1.00%
-1.50%
-2.00%
Mean Return
Negative Words
Positive Words
1
2
3
4
5
6
7
8
9
10
Weight Decile

Does document score convey information?

$$r_i = a + b \times \left(\sum_{j=1}^{J} \left(w_j F_{i,j} \right) \times \frac{1}{a_i} \right) + \varepsilon_i ,$$

We obtain the estimate of $\hat{w}_j$ that we use in Regression (8) using only data prior to the time that document i is made public. The null hypothesis is that our tone measure does not convey any incremental information to the market, in which case b would be zero, and the alternate hypothesis is $b > 0$.

Measurement Error

$$Score_i = \sum_{j=1}^{J} \left(w_j F_{i,j} \right) \times \frac{1}{a_i}, \qquad (4)$$

The measurement error in WP weights “diversified” away when we Compute the score.

Panel A: Positive Words

Overall

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term Weighting Scheme						
WP	0.429		0.387	0.254		0.247
	(2.54)		(2.60)	(2.95)		(2.96)
.idf		-0.092	-0.224		-0.035	-0.090
		(-1.86)	(-1.53)		(-1.78)	(-1.22)
Control Variables						
Size				-0.057	-0.025	-0.047
				(-0.68)	(-0.71)	(-0.54)
BM				-0.457	0.004	0.234
				(-0.04)	(0.07)	(0.02)
Volatility				-0.363	-0.772	-0.335
				(-1.97)	(-2.63)	(-2.01)
Turnover				-0.109	-0.078	-0.106
				(-1.23)	(-1.04)	(-1.22)
EAD-Ret				0.640	0.616	0.638
				(5.51)	(5.52)	(5.47)
Accruals				-0.282	-0.877	-0.287
				(-1.58)	(-1.79)	(-1.61)

Panel B: Negative Words

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term Weighting Scheme						
WP	0.429		0.396	0.268		0.263
	(2.47)		(2.59)	(2.50)		(2.58)
idf		-0.321	-0.313		-0.106	-0.120
		(-1.82)	(-1.77)		(-1.39)	(-1.62)
Control Variables						
Size				-0.069	-0.058	-0.052
				(-0.84)	(-0.69)	(-0.62)
BM				1.670	8.074	2.256
				(0.14)	(0.64)	(0.19)
Volatility				-0.434	-0.472	-0.410
				(-2.27)	(-2.42)	(-2.27)
Turnover				-0.118	-0.136	-0.114
				(-1.34)	(-1.50)	(-1.33)
EAD-Ret				0.641	0.607	0.638
				(5.54)	(5.86)	(5.50)
Accruals				-0.280	-0.283	-0.286
				(-1.61)	(-2.07)	(-1.64)

Panel C: Both Positive and Negative Scores

Rank Correlation of Positive and Negative Scores=0.3452

	Models	
	(7)	(8)
Term Weighting Scheme		
WP (Positive Words)	0.371	0.230
	(2.54)	(2.96)
WP (Negative Words)	0.274	0.211
	(2.33)	(2.09)
Control Variables		
Size		-0.066
		(-0.81)
BM		-1.727
		(-0.16)
Volatility		-0.345
		(-1.91)
Turnover		-0.100
		(-1.19)
EAD-Ret		0.641
		(5.54)
Accruals		-0.278
		(-1.59)

Choice of Lexicon

- Inclusion of irrelevant words
  - Use Harvard List
- Incomplete lexicon
  - Randomly exclude 50% of the words from the LM lexicon

Panel A: Additional Word Lists

	Hvd-Pos	Hvd-Neg	Pos_Omit	Neg_Omit
Term Weighting Scheme				
WP	0.196	0.409	0.296	0.253
	(2.13)	(2.51)	(3.75)	(4.00)
Control Variables				
Size	-0.062	-0.064	-0.044	-0.062
	(-0.74)	(-0.77)	(-0.52)	(-0.74)
BM	2.249	2.861	-0.207	1.464
	(0.19)	(0.23)	(-0.02)	(0.12)
Volatility	-0.430	-0.449	-0.370	-0.435
	(-2.26)	(-2.31)	(-1.98)	(-2.22)
Turnover	-0.120	-0.123	-0.096	-0.114
	(-1.33)	(-1.36)	(-1.15)	(-1.27)
EAD-Ret	0.639	0.640	0.640	0.639
	(5.48)	(5.51)	(5.50)	(5.47)
Accruals	-0.284	-0.288	-0.298	-0.276
	(-1.59)	(-1.63)	(-1.71)	(-1.55)

Panel B: Differences in Slope Coefficients between Hvd/Omitted lists and Pos/Neg lists

	(Positive-Hvd_Pos)	(Negative-Hvd_Neg)	(Positive-Pos_Omit)	(Negative-Neg_Omit)
ΔWP	0.058	-0.155	-0.040	-0.030
	(0.81)	(-0.49)	(-1.32)	(-0.62)

Timeliness of Market Reaction

Panel A: Positive Words

	Event Windows		
	+5 to +9	+5 to +14	+5 to +26
Dependent Variable			
Market-adjusted returns	0.143	0.330	0.440
	(2.59)	(1.93)	(0.64)
Size-adjusted returns	0.169	0.358	0.441
	(2.14)	(1.99)	(0.42)

Panel B: Negative Words

	Event Windows		
	+5 to +9	+5 to +14	+5 to 26
Dependent Variable			
Market-adjusted returns	0.089	0.176	0.323
	(2.11)	(1.71)	(0.87)
Size-adjusted returns	0.081	0.142	0.331
	(1.92)	(1.88)	(0.56)

Conclusion

- WP term-weighted document score captures document tone more reliably than other approaches in the literature
- WP term-weighting scheme reliably measures positive tone where other approaches had limited success
- Term weights at least as important as choice of lexicon
- Market slow to react to the tone of 10-Ks